NLP-Based Fusion Approach to Robust Image Captioning

Riccardo Ricci, *Graduate Student Member, IEEE*, Farid Melgani, *Fellow, IEEE*, José Marcato Junior, *Member, IEEE*, and Wesley Nunes Gonçalves

***Abstract*—Robustness in remote sensing image captioning is crucial for real-world applications. However, most of the research focuses on improving the performance of single captioning algorithms, either by introducing novel feature processing units or metatasks that indirectly improve the captioning performance. Despite indisputable improvements in performance, we argue that relying on the output of a single model can be critical, especially when data scarcity limits the generalization capability of the trained algorithms. Focusing on the advantages of ensembles for improving robustness, we propose different ways to select or generate a single most coherent caption from a set of predictions made by different captioning algorithms. The disjunction between the two phases of prediction and selection/generation provides high flexibility for inserting different captioning algorithms, each with its peculiarities and strengths. In this context, based on neural natural language processing tools, our approach can be considered as an additional fusion block that enables higher robustness with a contained complexity burden.**

***Index Terms*—Bidirectional encoder representations from transformers (BERT), contrastive language-image pretraining (CLIP), ensemble fusion, generative pretrained transformer (GPT), image captioning, natural language processing (NLP).**

I. Introduction

REMOTE sensing imagery is one of the biggest sources of data about our planet. Every day, thousands of terabytes of visual data are sent to Earth from various satellites orbiting the globe. The information within these images can fuel many applications, but the rate at which it must be analyzed makes it unfeasible to rely on manual approaches. Thus, automatic techniques able to discover, explain, and analyze information hidden in remote sensing scenes are becoming increasingly important in the remote sensing landscape. Different trends within the remote sensing community are pushing toward this goal. Some studies focus solely on the visual aspect, tackling tasks such as scene classification [1], [2], semantic segmentation [3], [4], and object detection [5], among others.

Following the recent interest in visual-language models, another line of work tries to connect visual and language modalities, enhancing the interaction between the user and the machine that analyzes the image through natural language. This joint interaction between visual information and natural language empowers different applications. Some widely studied applications of vision-language models in remote sensing are presented here. Text-based image retrieval [6], [7] deals with retrieving the most coherent image satisfying a textual description. A second example is text-based image generation [8], [9], which focuses on the generation of a remote sensing image given a purely textual description of it. Visual question-answering [10] studies the possibility of conferring more interactive abilities to a vision-language model, enabling it to understand and answer user queries about the image contents. A last example is image captioning [11], which shares architectural similarities with visual question answering, but whose goal is to generate a single descriptive sentence of the image contents rather than answering questions interactively.

Image captioning has been tackled extensively in the computer vision community, focusing on natural images. The majority of approaches employ the encoder–decoder structure, coupled with the attention mechanism. For example, He et al. [12] exploit the transformer architecture focusing on the modeling of the visual scene by decomposing image regions into parent, neighbor, and child. Then, different subattention modules are used to model the interactions between the different spatial and semantic regions of the image, improving the feature extraction ability of the network, and in turn, the captioning performance. Yang et al. [13] employ a joint attribute prediction submodule and semantic rewards to guide the network in outputting more informed and detailed descriptions. Further works such as [14] explore vision-language pretraining as a first step toward image captioning. In this paradigm, the network is first pretrained on a large corpus of data using self-supervised objectives, and then, finetuned using labeled data, achieving stronger performance compared to networks trained solely on the labeled data.

In our work, we focus on remote sensing image captioning (RSIC), which is attracting attention due to its several implications and real-life use cases. An algorithm that can automatically describe a remotely sensed (RS) picture can empower tasks such

Manuscript received 10 November 2023; revised 22 March 2024 and 2 May 2024; accepted 5 June 2024. Date of publication 12 June 2024; date of current version 1 July 2024. This work was supported in part by a project entitled "Deep Learning for PrecisionFarming Mapping," cofounded by the Italian Ministry of Foreign Affairs, International Cooperation and the Brazilian National Council of State Funding Agencies and Fundect (p. 71/001.902/2022), and in part by CAPES PrInt (p: 88881.311850/2018-01). *(Corresponding author: Farid Melgani.)*

Riccardo Ricci and Farid Melgani are with the Department of Information Engineering and Computer Science, University of Trento, 38123, Trento, Italy (e-mail: riccardo.ricci-1@unitn.it; melgani@disi.unitn.it).

José Marcato Junior and Wesley Nunes Gonçalves are with the Federal University of Mato Grosso do Sul, Cidade Universitária, Campo Grande, MS 79070-900, Brazil (e-mail: jose.marcato@ufms.br; wesley.goncalves@ufms.br).

Digital Object Identifier 10.1109/JSTARS.2024.3413323

as image retrieval [6] and decrease the expertise required to analyze an RS image by translating the information within into a more accessible and general format. In the remote sensing community, the task of image captioning lags behind traditional natural image captioning [15], mainly because of the scarcity of large-scale datasets and the intrinsic complexity of remote sensing scenes. To date, the main sources of data for remote sensing image captioning are three datasets, namely Sidney-Captions [16], UCM-Captions [16], and RSICD [17]. Due to the complexity and cost of manually labeling RS images for captioning, remote sensing image captioning datasets are orders of magnitude smaller than the natural image counterparts [18], [19], forcing remote sensing researchers to adopt all kinds of tricks to improve the performance of their algorithms. However, we argue that a major concern is the generalization capability. It is fair to assume that training an algorithm on a small amount of samples can cause errors and unwanted behaviors when applied on slightly out-of-distribution samples. This is further exacerbated in image captioning, where data points lie in hyperdimensional spaces and other problems exist, such as error accumulation during inference. Most of the literature, here not restricting the analysis to remote sensing, has been focused on improving captioning performance by introducing several variations to the vanilla pipeline, proposed first in the remote sensing field by Vinyals et al. [15]. Examples are the Bahdanau attention mechanism [20], introduced for text translation, and then, adopted for captioning, the extraction of multilevel features from images that enable more fine-grained cross-modal interaction between text and image features [21], the insertion of metatasks during training to aid the learning of better image and text representations [22], to the more recent transformer architecture [23], which relies solely on the attention mechanism, exhibiting promising performance in remote sensing image captioning [24]. Despite the success of these methods in enhancing remote sensing image captioning performance, we argue that there are no guarantees that such improvements will hold in more realistic scenarios. This limits the confidence in building applications based on these algorithms, given that the reliability of these methods is still unclear. Ensembles have been studied extensively in the literature to solve challenging classification problems and have been shown to boost both accuracy and robustness [25]. In [26], for example, an ensemble is adopted to boost the accuracy of change detection maps. Another example is in [27], where Melgani and Bazi apply ensembles to boost multiclass classification accuracy. Despite the success of the application of ensembles in classification problems, this line of research received no attention in remote sensing image captioning literature. Katpally and Bansal [28] apply the ensemble in a natural image captioning scenario, combining the output probabilities of each captioner, and sampling the next word upon the resulting aggregated probability. The authors show in the paper that this approach led to a slight boost in captioning performance; however, we argue that approaching an ensemble for image captioning in such a way poses significant constraints. First, every captioner must be trained using the same dictionary of words, limiting the exploration capability and the inventiveness of different captioners trained with different vocabularies. Second, the majority of the captioners must be highly targeted to the application domain, at the risk of severely impacting the performance. In this article, we propose a different paradigm. The idea is to leverage the ensemble concept a posteriori, when all the models have already created their output description. In this way, there are no predefined constraints, leaving the freedom to adopt different captioning architectures, adopt different caption generation schemes, and use different vocabularies. We propose three ways to leverage the ensemble of captioners, each with strengths and drawbacks, providing a thorough analysis of each of them on different scenarios. To summarize, the main contributions of this article are the following.

1) We propose three strategies to employ a postgeneration ensemble of different architectures in the context of image captioning.
2) We validate the ensemble on four datasets, three remote sensing image captioning datasets, and a UAV image captioning dataset.
3) We validate the robustness and performance of the ensemble in different use cases, simulating different operative conditions.

The rest of this article is organized as follows. Section II provides a brief introduction to image captioning, followed by an in-depth explanation of our three proposed ensemble fusion strategies. Section III briefly describes the datasets employed in this study. Section IV is dedicated to the experimental validation of the proposed methodologies, encompassing both quantitative and qualitative results, along with discussions on the behavior of the ensemble techniques under different scenarios. In Section V, we elaborate on some general insights and implications drawn from the study. Finally, Section VI concludes this article.

II. Methodology

The complete pipeline of our proposed method is represented in Fig. 1. The two main stages are caption generation stage and postgeneration fusion stage. The caption generation stage employs a set of different algorithms, each of them producing a caption for the input image. The postgeneration fusion stage employs an ensemble module that ingests the set of generated captions (and optionally the image) as input and produces a single output best caption. The following subsections explore in detail our choices for both stages.

A. Caption Generation Stage

As stated in [25], the most important requirement for an ensemble is that each participant is different compared to the others. In the context of our proposal and considering that we are applying the ensemble a posteriori, differences can derive from training strategies, training sets, architectures, or vocabularies. For the sake of this study, and without lack of generalization, we generate an ensemble of $M = 7$ different captioners by employing a combination of architectural diversity and training set diversity. Specifically, we built four captioners following [15], which defines a captioner architecture as composed of two blocks: encoder and decoder. We use different combinations, as

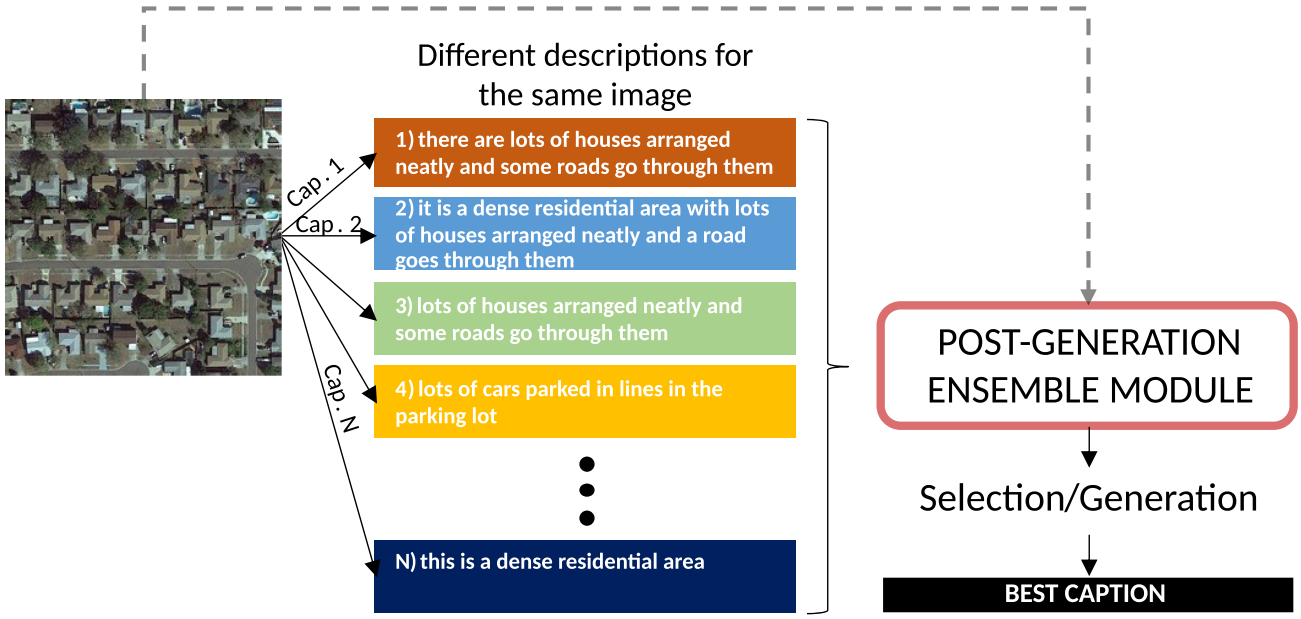


Fig. 1. Conceptual pipeline of our proposed captioner ensemble fusion approach. From an image, several captions are generated using different captioners. A postgeneration ensemble module takes this set of captions and optionally the image, outputting a single best caption.

TABLE I
ALGORITHMS INCLUDED IN OUR ENSEMBLE PROPOSAL

	Encoder	Decoder
CC-a	VGG-16	GRU
CC-b	ResNet-50	GRU
CC-c	ResNet-50	Transformer
CC-d	ViT	Transformer
MLAT [29]	ResNet-50	Transformer
Blip-2 [30]	ViT	FlanT5 XXL [32]
CapDec [31]	RN-50x4	GPT-2

CC stands for custom captioner.

depicted in Table I, and further included three more captioners from the literature, namely MLAT [29], Blip-2 [30], and CapDec [31]. It is worth noting that CapDec and Blip-2 are not specifically tailored to remote sensing scenes, but are trained on natural images.We start by formulating the RS image captioning task considering a set of N images $I = [I_1, \ldots, I_N]$, where I_i is the ith image in the set. Let $C_i = [c_{i,j}]_{j=1}^{K_i}$ be the caption set associated with the ith image, where $c_{i,j}$ is the jth caption of the set and K_i is the number of captions associated with the ith image. Each caption can be seen as a set of L ordered tokens $c_{i,j} = [t_k]_{k=1}^{L}$, where L is the length of the caption in tokens.

Image captioning aims to accurately estimate the conditional probability of the subsequent token t_t, given both the image and the sequence of preceding tokens, depicted as follows:

$$p(t_t|t_{1:t-1}, I). \tag{1}$$

In all the models employed in our ensemble, this conditional probability is estimated using neural network architectures, even though this is not inherently a restriction within the scope of our proposed framework. The image I acts as a conditioning variable and its salient features are extracted using the encoder. For this purpose, we implemented two distinct encoding strategies: a convolutional neural network (CNN) and a vision transformer (ViT). Simultaneously, the conditioning based on previous words was accomplished through autoregressive networks. Our experimentation spanned two variations, namely the gated recurrent unit (GRU) and a transformer decoder (TD).

While a comprehensive review of the inner mechanisms of our chosen encoders and decoders remains vital, the subsequent sections focus on the nuances of the postprocessing block. For detailed descriptions of how image captioning is achieved from a mathematical viewpoint, we suggest some informative papers such as [15] and [24] and surveys such as [33].

B. Postgeneration Fusion Stage

Within this block, we have envisioned two primary postprocessing methodologies: selection and generation. Broadly, selection proves advantageous in scenarios marked by pronounced uncertainty among captioning models, where a significant proportion of generated captions might lack relevance to the actual image content. Conversely, generation becomes particularly pertinent when there is substantial semantic concordance among captioners, but discrepancies or inaccuracies in the generated captions could compromise both their performance and legibility.

In detail, we introduce two selective paradigms: the naïve selection and the contrastive language-image pretraining (CLIP)-coherence selection. Complementarily, we also propose a generative approach grounded in the variational autoencoder (VaE) framework.

1) Naïve Selection: Naïve selection, depicted in Fig. 2, is a text-only strategy, in which the selection process relies solely on

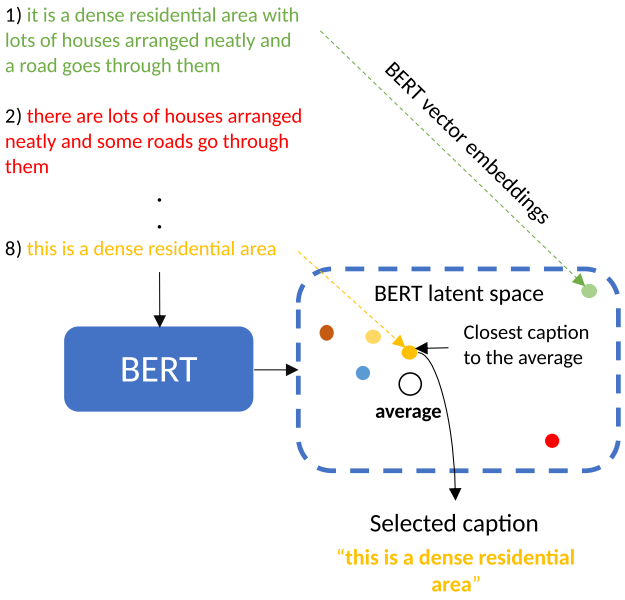


Fig. 2. Pipeline of the naïve selection strategy. Each caption generated in the first stage is projected into a semantic latent space using a BERT pretrained model. The average of the embeddings is used as a reference, and the caption whose embedding is closest to it is selected.

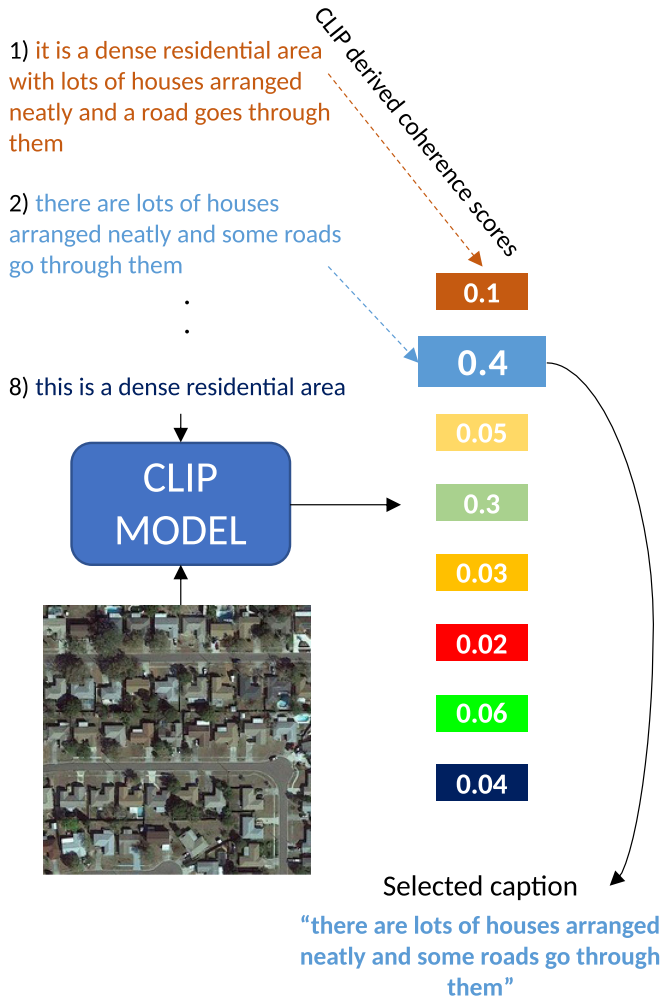


Fig. 3. Pipeline of the CLIP-coherence selection strategy. Using a pretrained CLIP model, an image-text coherence score is derived for each caption generated in the first stage. The caption showing the higher coherence with the given image is selected as the final output.

the generated set of captions. This strategy leverages a distilled bidirectional encoder representations from transformers (BERT [34]) model, particularly suited to extract semantic vector representations of sentences. Devlin et al. [34] showed that a massive pretraining allows the BERT model to extract representations of the input sentence that can benefit a wide suite of tasks such as question-answering and language inference. We adopted the distilled version from [35] to decrease the computational requirements of the entire pipeline. The distilled model has been finetuned with a contrastive learning paradigm in which the model is asked to predict which sentence in a batch is more likely to be paired with the input sentence. In this way, the network sharpens its knowledge about semantically related and unrelated sentences, reflecting this knowledge in its latent space, where representations are more compatible for related sentences and less compatible otherwise. The naïve selection strategy proceeds as follows.

1) *Embedding extraction:* Each caption is encoded in a vector embedding using the sentence transformer model [36], resulting in a set of vector embeddings $\{e_1, e_2, \ldots, e_N\}$.
2) *Aggregation:* The average of all the vector embeddings $\bar{v}$ is computed as $\bar{v} = \frac{1}{N}\sum_{i=1}^{N} e_i$. This average $\bar{v}$ serves as the prototype vector encapsulating the average semantic content of the set of captions.
3) *Selection:* The embedding falling closest to the prototype is identified, and the related caption is selected as the final result.

Within the landscape of ensemble techniques, this approach shares similarities to the majority voting principle, albeit recontextualized for caption generation.

2) CLIP-Coherence Selection: CLIP-coherence Selection employs the CLIP model [37], a multimodal framework that integrates both textual and visual data. The pipeline is represented in Fig. 3. The CLIP model consists of two branches: the textual branch incorporates a transformer encoder [23] and the visual branch utilizes a ViT [38]. Both branches generate vector representations for images and text, respectively. A contrastive learning paradigm is employed to learn a common feature space where consistent image–text pairs are projected in similar regions, while inconsistent pairs are projected far apart. This paradigm facilitates the encapsulation of intrinsic similarities between text and image data within a unified semantic space. Remarkably, the CLIP model has demonstrated robust capabilities in zero-shot classification and retrieval tasks in the natural scenario, further extending in the remote sensing field [39]. To evaluate the efficacy of the CLIP-Coherence Selection strategy, we conducted experiments using two versions of the CLIP model. The first version is the general-purpose CLIP model [40] as released by OpenAI, trained on a comprehensive corpus of web-scraped text–image pairs. The second [41] is a specialized variant that has undergone fine tuning for remote sensing applications. This version has displayed enhanced performance

metrics in remote sensing-related tasks, such as remote sensing image retrieval and remote sensing zero-shot classification. The CLIP-coherence selection procedure is structured as follows.

1) *Embedding extraction:* The image vector embedding and the vector embedding of each caption are extracted using respectively the vision branch and the textual branch of the CLIP model.
2) *Similarity computation:* A list of scores is generated by computing the similarity between the image vector embedding and each of the caption vector embedding.
3) *Caption selection:* The caption that registers the highest compatibility score is selected.

From our perspective, the CLIP-coherence selection strategy facilitates the use of an ensemble of *specialized* captioners. Each member of this ensemble can excel in a distinct subset of scenes, eliminating the need for a universal captioner across the entire spectrum of possible scenes. Notably, this procedure does not involve any averaging. Consequently, even in scenarios in which only a single caption exhibits coherence with the image, this strategy can, in principle, be able to isolate it, irrespective of the noise provided by the remaining candidates.

3) VaE Fusion Strategy: The VaE fusion is a text-only strategy based on the VaE framework. The pipeline is represented in Fig. 4. The idea is to condense the information of the set of generated captions and distill a single output caption. We decided to adopt the VaE framework instead of a plain autoencoder for its ability to learn a smooth latent-to-output distribution through noisy sampling in the latent space. It has been demonstrated that textual interpolation in the VaE latent space yields smoother and more plausible outputs [42] than standard autoencoders. Due to the inherent data restriction, instead of training a VaE from scratch, we decided to finetune a pretrained VaE-based language model called OPTIMUS [43] .

OPTIMUS is composed of two subnetworks: BERT [34] and GPT-2 [44]. As mentioned in the naïve strategy, BERT acts like an encoder, taking a sentence and producing its semantic embedding. In this approach, the embedding is further projected to a VaE latent space. During training, noisy sampling is used to obtain a noisy latent vector, that eventually conditions the decoder (GPT-2) to reconstruct the input sentence. OPTIMUS has been pretrained on roughly 2 M sentences from English Wikipedia. To make it more targeted to modeling sentences, the authors used preprocessing to isolate sentences of a maximum length of 64 (tokens). Starting from the pretrained weights, we produce four fine-tuned OPTIMUS models, one for each dataset, using captions in the respective trainsets. The VaE fusion strategy, as illustrated in Fig. 4, entails the following steps.

1) *Projection:* Each caption is mapped to its latent space representation (embedding) using the encoder (BERT), resulting in a set of embeddings $\{e_1, e_2, \ldots, e_N\}$.
2) *Aggregation:* The average of all the embeddings is computed as $\bar{v} = \frac{1}{N}\sum_{i=1}^{N} e_i$
3) *Decoding:* Conditioned on $\bar{v}$, a new caption is decoded using the decoder (GPT-2).

The idea is to leverage the modeling capability of the variational latent space to retain a condensed representation of the input captions, thus discarding noise from possible errors or misspellings. Decoding from such a representation can be seen as extracting the condensed semantic meaning of the input set of captions. We formulate this strategy as a way to deal with possible syntactic errors in the input captions. Indeed, such a scenario cannot be tackled using selective strategies, which are restricted to the already generated candidates. Details of architecture and training are provided in Section V.

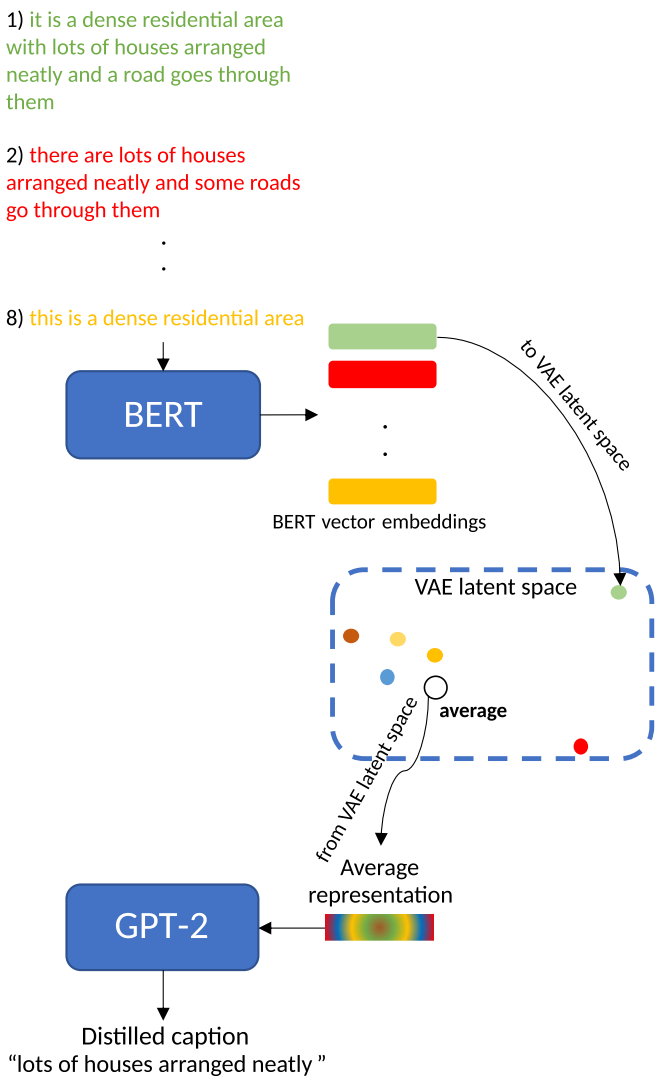


Fig. 4. Pipeline of the VaE fusion strategy. The embeddings derived using BERT are further projected in a smooth VaE latent space. There, the average representation is computed, and used to condition a GPT-2-based decoder to distill the final caption.

III. Dataset

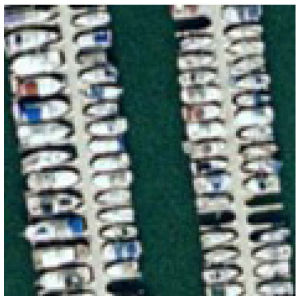
The testbed of this work is composed of four datasets. Three have been widely adopted in remote sensing image captioning literature: UCM-Captions, SIDNEY-Captions, and RSICD dataset. In addition, we decided to include another dataset for UAV imagery image captioning, called UAV-Captions [45], to further increase the diversity of the testing set. Examples of images and captions can be found in Fig. 5. The next subsections explore in detail each dataset.

1) this square playground is made up of two tennis courts coloured with green
2) two small courts locate in a large grass
3) two tennis courts surrounded by green meadows are near to several buildings
4) two tennis fields are surround by green meadows
5) two tennis courts are surrounded by a large piece of green meadow

(a)

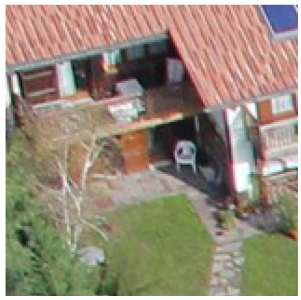
1) a residential area with many houses divided into rectangles by some roads
2) there are many houses densely arranged and divided into rectangles by some roads
3) a residential area with lots of houses arranged neatly
4) lots of houses arranged neatly and divided into rectangles by some roads
5) there are lots of houses arranged neatly in the dense residential area

(b)



1) lots of boats docked in lines at the harbor
2) lots of boats docked neatly at the harbor
3) many boats docked neatly at the harbor and the water is deep blue
4) many boats docked neatly at the harbor and just a few positions are free
5) lots of boats docked in lines at the harbor and just a few positions are free

(c)



1) grass field with building facade and red roof on the top
2) red roof on top and building facade near grass field with tree on the bottom left
3) one large tree at left is next to red roof at upper to building facade and to grass field

(d)

Fig. 5. Examples from (a) RSICD, (b) Sidney-Captions, (c) UCM-Captions, and (d) UAV-Captions datasets.

A. UCM-Captions Dataset

The UCM-Captions dataset includes 2100 aerial images, each of which has a size of 256 × 256 pixels with a spatial resolution of one foot. This dataset is defined based on the UC Merced Land Use dataset, in which each image is associated with one of 21 land-use classes. Each image in the UCM-Captions dataset was annotated with five captions, resulting in 10 500 captions. Although five captions per image are considered, captions belonging to the same classes are very similar in both datasets. The Sydney-Captions and the UCM-Captions datasets were initially built for scene classification problems with a few images.

B. SIDNEY-Captions Dataset

The Sydney-Captions dataset includes 613 images, each of which has a size of 500 × 500 pixels with a spatial resolution of 0.5 m. This dataset was built based on the Sydney scene classification dataset, which includes RS images annotated with one of the seven land-use classes. Each image in the Sydney-Captions dataset was annotated by the five captions, providing 3 065 captions in total.

C. RSICD Dataset

This dataset includes 10 921 images of size 224 × 224 with varying spatial resolutions. Each image is described with five captions. For some images, caption replication has been used to artificially augment the number of captions associated with the image up to reach the predefined number of five captions per image. The total number of captions in the dataset is 54 605.

D. UAV-Captions Dataset

The UAV-Captions dataset was acquired near the city of Civezzano, Italy, at different off-nadir angles on October 17, 2012. The data acquisition was performed using a Canon EOS 550D camera, characterized by a CMOS APS-C sensor with 18 megapixels. The images are characterized by three channels (RGB) with a spatial resolution of approximately 2 cm. The obtained images are of size 5184 × 3456 pixels with 8-bit radiometric resolution. This dataset is composed of ten images that are subdivided into training (6), validation, and test. All the images are subdivided into a nonoverlapping grid of equal tiles of size 256 × 256 pixels. More in detail, 1746, 294, and 882 tiles are extracted from training, validation, and test images, respectively. Three different captions are manually produced for each tile.

IV. Experimental Validation

In our experimental setup, all custom-trained captioners uniformly employed the same scheme and hyperparameters for both training and inference. Each captioner utilized the tokenizer from BERT [34] with an embedding dimension of $d = 256$. For training, we used a batch size of $b = 8$, a learning rate of $\alpha = 1 \times 10^{-4}$, and a dropout probability of $p = 0.15$. We decided to freeze the parameters of the encoders, while keep updating the parameters of the decoders. To mitigate potential overfitting due to the limited dataset size, we implemented early stopping, monitoring the validation BLEU-4 metric. The AdamW optimizer was employed with a weight decay of $\lambda = 1 \times 10^{-7}$. All models have been developed using Pytorch, and trained on an NVIDIA Geforce RTX 3090 GPU. Notably, the VaE was finetuned for $e = 10$ epochs, adhering to the configuration for the base model's pretraining used by the original authors [43] .

To provide a comprehensive assessment of the proposed ensemble across diverse contexts, we delineated three evaluation scenarios. These scenarios are crafted to unveil important insights into the ensemble's performance under different conditions. The performance is reported using several common

TABLE II
UCM-CAPTIONS: STANDARD EVALUATION BLEU1-4 (B1-4), ROUGE (R), METEOR (M), AND CIDER (C)

	B1	*B2*	*B3*	*B4*	*R*	*M*	*C*
CC-a	0.79	0.72	0.66	0.61	0.74	0.40	3.03
CC-b	0.77	0.70	0.64	0.59	0.71	0.39	2.99
CC-c	0.54	0.45	0.38	0.34	0.48	0.22	1.56
CC-d	**0.82**	0.75	**0.70**	**0.65**	**0.77**	**0.43**	**3.19**
MLAT [29]	0.42	0.23	0.13	0.09	0.31	0.14	0.54
Blip-2 [30]	0.35	0.19	0.10	0.04	0.27	0.13	0.33
CapDec [31]	0.30	0.16	0.08	0.04	0.26	0.11	0.18
Naïve	0.80	0.72	0.66	0.60	0.74	0.40	3.09
CLIP-rsicdv2	0.69	0.61	0.56	0.52	0.62	0.34	2.49
CLIP-vitlarge14	0.55	0.44	0.38	0.34	0.45	0.24	1.48
VaE	0.76	0.67	0.60	0.54	0.70	0.37	2.69

Bold entries highlight the best results.

TABLE III
UAV-CAPTIONS: STANDARD EVALUATION BLEU1-4 (B1-4), ROUGE (R), METEOR (M), AND CIDER (C)

	B1	*B2*	*B3*	*B4*	*R*	*M*	*C*
CC-a	0.69	0.59	0.49	0.40	0.70	0.34	3.92
CC-b	0.70	0.59	0.48	0.38	0.69	0.34	3.76
CC-c	0.59	0.48	0.39	0.31	0.60	0.27	3.08
CC-d	0.68	0.57	0.46	0.37	0.69	**0.35**	3.56
MLAT [29]	0.13	0.03	0.01	0.00	0.13	0.06	0.04
Blip-2 [30]	0.21	0.07	0.03	0.00	0.18	0.10	0.11
CapDec [31]	0.13	0.05	0.01	0.00	0.13	0.09	0.04
Naïve	**0.73**	**0.62**	**0.52**	**0.42**	**0.72**	**0.35**	**4.00**
CLIP-rsicdv2	0.35	0.24	0.18	0.13	0.35	0.18	1.08
CLIP-vitlarge14	0.27	0.16	0.11	0.07	0.25	0.16	0.52
VaE	0.68	0.57	0.45	0.35	0.66	0.31	3.28

Bold entries highlight the best results.

TABLE IV
SIDNEY-CAPTIONS: STANDARD EVALUATION BLEU1-4 (B1-4), ROUGE (R), METEOR (M), AND CIDER (C)

	B1	*B2*	*B3*	*B4*	*R*	*M*	*C*
CC-a	**0.77**	**0.68**	**0.61**	**0.55**	**0.70**	**0.38**	**2.35**
CC-b	0.73	0.63	0.56	0.49	0.66	0.35	2.11
CC-c	0.69	0.60	0.52	0.45	0.64	0.34	1.91
CC-d	0.76	0.67	0.60	0.53	**0.70**	**0.38**	2.31
MLAT [29]	0.47	0.24	0.13	0.07	0.29	0.15	0.24
Blip-2 [30]	0.33	0.18	0.12	0.07	0.28	0.11	0.10
CapDec [31]	0.31	0.11	0.00	0.00	0.26	0.08	0.08
Naïve	0.76	0.67	0.60	0.53	**0.70**	**0.38**	2.34
CLIP-rsicdv2	0.59	0.46	0.39	0.34	0.50	0.25	1.58
CLIP-vitlarge14	0.56	0.44	0.37	0.33	0.46	0.23	1.11
VaE	0.71	0.61	0.54	0.47	0.65	0.35	1.60

Bold entries highlight the best results.

metrics for image captioning: BLEU1-4 (B1-4) [46], Rouge (R) [47], Meteor (M) [48], and Cider (C) [49].

In tables, a color encoding scheme is employed for ease of visualization. Specifically, the colors are utilized as follows.

1) LightCyan : Algorithm trained on UCM-Captions.
2) LightRed : Algorithm trained on RSICD-Captions.
3) LightGreen : Algorithm trained on SIDNEY-Captions.
4) LightYellow : Algorithm trained on UAV-Captions.

TABLE V
RSICD-CAPTIONS: STANDARD EVALUATION BLEU1-4 (B1-4), ROUGE (R), METEOR (M), AND CIDER (C)

	B1	*B2*	*B3*	*B4*	*R*	*M*	*C*
CC-a	0.61	0.44	0.34	0.27	0.46	0.24	0.71
CC-b	0.60	0.43	0.32	0.26	0.44	0.24	0.69
CC-c	0.54	0.36	0.26	0.20	0.38	0.20	0.48
CC-d	0.63	0.47	0.37	0.30	0.48	0.26	0.81
MLAT [29]	**0.65**	**0.49**	**0.39**	**0.32**	0.49	**0.27**	**0.90**
Blip-2 [30]	0.34	0.16	0.08	0.04	0.24	0.10	0.20
CapDec [31]	0.35	0.16	0.07	0.03	0.23	0.10	0.13
Naïve	**0.65**	**0.49**	**0.39**	**0.32**	**0.50**	**0.27**	0.86
CLIP-rsicdv2	0.58	0.42	0.32	0.26	0.43	0.23	0.73
CLIP-vitlarge14	0.46	0.28	0.19	0.15	0.31	0.15	0.40
VaE	0.63	0.44	0.33	0.26	0.46	0.24	0.73

Bold entries highlight the best results.

TABLE VI
RSICD-CAPTIONS: GENERALIZATION EVALUATION BLEU1-4 (B1-4), ROUGE (R), METEOR (M), AND CIDER (C)

	B1	*B2*	*B3*	*B4*	*R*	*M*	*C*
CC-a	0.35	0.16	0.08	0.04	0.23	**0.10**	0.12
CC-b	0.37	0.17	**0.09**	0.04	**0.24**	**0.10**	0.13
CC-c	0.33	0.14	0.06	0.02	0.21	0.08	0.07
CC-d	0.38	**0.18**	**0.09**	**0.05**	**0.24**	**0.10**	0.14
CC-a	0.31	0.09	0.03	0.01	0.20	0.09	0.04
CC-b	0.31	0.09	0.03	0.01	0.20	0.09	0.05
CC-c	0.32	0.12	0.04	0.01	0.21	0.09	0.05
CC-d	0.33	0.13	0.05	0.02	0.21	**0.10**	0.08
CC-a	0.19	0.08	0.02	0.00	0.17	0.06	0.03
CC-b	0.20	0.09	0.02	0.00	0.17	0.06	0.03
CC-c	0.16	0.07	0.02	0.00	0.16	0.05	0.03
CC-d	0.20	0.09	0.02	0.00	0.18	0.05	0.02
Blip-2 [30]	0.33	0.16	0.07	0.04	0.23	**0.10**	0.19
CapDec [31]	0.35	0.15	0.07	0.03	0.23	**0.10**	0.11
Naïve	0.36	0.16	0.07	0.03	0.22	**0.10**	0.14
CLIP	0.35	0.17	0.08	0.04	**0.24**	**0.10**	**0.21**
VaE	**0.40**	0.17	0.08	0.03	**0.24**	**0.10**	0.10

Bold entries highlight the best results.

A. *Scenario 1: Standard Evaluation*

In this experiment, we exclusively use custom captioners that are trained and evaluated on the same dataset. The ensemble thus comprises our four custom-trained captioners in conjunction with three pre-existing captioners from the literature, for a total of $N = 7$ captioners. The performance is reported for each of the four datasets. Such an approach is the canonical scenario frequently observed in remote sensing image captioning literature. In each table, the first seven rows represent the performance of the single captioners, while the last four results of the ensemble using the various strategies .

The ensemble's performance is summarized in Tables II–V. Most custom-built captioners achieve satisfactory results, with the notable exceptions of Blip-2 and CapDec, which lag behind across all evaluation metrics. It is important to highlight that these models have not been specifically trained on remote sensing data, making their lower scores somewhat expected. A critical issue arises when considering the limitations of existing evaluation metrics. For example, if the ground-truth caption

TABLE VII
SIDNEY-CAPTIONS: GENERALIZATION EVALUATION BLEU1-4 (B1-4), ROUGE (R), METEOR (M), CIDER (C)

	B1	*B2*	*B3*	*B4*	*R*	*M*	*C*
CC-a	0.48	0.36	0.26	0.20	0.38	0.17	0.38
CC-b	0.38	0.25	0.12	0.07	0.33	0.13	0.14
CC-c	0.32	0.20	0.10	0.00	0.25	0.09	0.10
CC-d	0.53	0.41	0.32	**0.27**	0.42	0.20	0.44
CC-a	0.41	0.16	0.09	0.05	0.26	0.13	0.16
CC-b	0.43	0.18	0.09	0.05	0.26	0.13	0.21
CC-c	0.33	0.09	0.03	0.00	0.21	0.11	0.09
CC-d	0.47	0.22	0.12	0.07	0.29	0.16	0.27
CC-a	0.12	0.05	0.00	0.00	0.14	0.05	0.02
CC-b	0.16	0.08	0.00	0.00	0.14	0.04	0.04
CC-c	0.15	0.07	0.00	0.00	0.16	0.04	0.02
CC-d	0.15	0.08	0.00	0.00	0.16	0.05	0.02
MLAT [29]	0.49	0.25	0.13	0.08	0.30	0.16	0.28
Blip-2 [30]	0.31	0.18	0.12	0.00	0.28	0.10	0.10
CapDec [31]	0.28	0.10	0.04	0.00	0.22	0.07	0.07
Naïve	0.49	0.26	0.14	0.08	0.32	0.16	0.25
CLIP	0.41	0.24	0.15	0.06	0.31	0.14	0.27
VaE	**0.64**	**0.46**	**0.33**	0.24	**0.47**	**0.24**	**0.86**

Bold entries highlight the best results.

TABLE VIII
UAV-CAPTIONS: GENERALIZATION EVALUATION BLEU1-4 (B1-4), ROUGE (R), METEOR (M), AND CIDER (C)

	B1	*B2*	*B3*	*B4*	*R*	*M*	*C*
CC-a	0.17	0.06	0.02	0.00	0.17	0.07	0.04
CC-b	0.19	0.08	0.02	0.00	0.20	0.07	0.04
CC-c	0.21	0.12	0.03	0.01	0.23	0.08	0.05
CC-d	0.18	0.07	0.02	0.01	0.18	0.07	0.05
CC-a	0.10	0.02	0.00	0.00	0.09	0.05	0.04
CC-b	0.11	0.03	0.01	0.00	0.10	0.05	0.04
CC-c	0.10	0.01	0.00	0.00	0.09	0.04	0.04
CC-d	0.10	0.03	0.01	0.01	0.09	0.05	0.04
CC-a	0.10	0.03	0.01	0.00	0.09	0.04	0.02
CC-b	0.13	0.03	0.00	0.00	0.12	0.05	0.03
CC-c	0.17	0.07	0.02	0.00	0.17	0.06	0.04
CC-d	0.19	0.09	0.03	0.00	0.20	0.07	0.05
MLAT [29]	0.12	0.02	0.00	0.00	0.13	0.05	0.04
Blip-2 [30]	0.20	0.08	0.03	0.02	0.19	0.11	0.11
CapDec [31]	0.14	0.06	0.02	0.01	0.13	0.09	0.05
Naïve	0.14	0.05	0.01	0.00	0.14	0.06	0.05
CLIP	0.19	0.08	0.03	0.02	0.18	0.10	0.10
VaE	**0.37**	**0.24**	**0.15**	**0.09**	**0.34**	**0.15**	**0.37**

Bold entries highlight the best results.

states "This is a part of a city with houses arranged in lines," and the prediction is "That is a view of a residential area with many houses arranged neatly," conventional metrics will assign a low score, despite the high degree of semantic similarity between the two captions. Despite these challenges, the ensemble method performs consistently well, often aligning with or exceeding the best results from individual captioners (see Table III). The naïve approach, despite its simplicity, proves to be a competitive baseline, occasionally surpassing the performance of single, specialized models. Moreover, the VaE framework demonstrates its utility by effectively capturing the essence of the set of captions and distilling this information into a single, informative

CC-a - this is a beach with blue sea and white sands	CC-a - it is a straight runway with some mark lines on it
CC-b - the waves slapping a white sand beach	CC-b - there are two straight freeways closed to each other
CC-c - this is a dense forest with green waters and grass	CC-c - there are some mark lines on the straight runway
CC-d - the waves slapping a white sand beach over and over again	CC-d - there are four airplanes in the airport
MLAT - a white waves is near a yellow beach	MLAT - many planes are parked in an airport
Blip2 - a man is walking on the beach	Blip2 - aerial view of a small airport
CapDec - A woman in a white dress is standing in the snow.	CapDec - A group of planes parked in an airport.
Naive - a white waves is near a yellow beach	Naive - A group of planes parked in an airport.
CLIP-rsicdv2 - the waves slapping a white sand beach over and over again	CLIP-rsicdv2 - A group of planes parked in an airport.
CLIP-vitlarge14 - a man is walking on the beach	CLIP-vitlarge14 - A group of planes parked in an airport.
VaE - this is a beach with white sands and blue waters	VaE - there are two airplanes with black fuselage taxiing on the runway

Fig. 6. Qualitative results on UCM-Captions, scenario 1.

output. Unexpectedly, the CLIP selection pipeline appears to hinder performance, particularly the non finetuned variant. Two factors may contribute to this outcome. First, the CLIP model is not particularly specialized for remote sensing image captioning, and second, it often selects captions produced by Blip-2, which are generally coherent, but arranged using a different word distribution, and consequently, receive lower evaluation scores. These observations indicate a need to reconsider the evaluation metrics used in image captioning. Current metrics, which prioritize syntax and exact word matching over semantic integrity, may not fully capture the quality of a generated caption. This observation is confirmed in the qualitative results depicted in Figs. 6 and 7. It can be noticed how the ensemble, and especially the selective strategies can often select a very coherent caption.

B. Scenario 2: Generalization Evaluation

This scenario is designed to simulate a more real-world operational environment characterized by a higher diversity between training and testing data. Specifically, each dataset, in turn, undergoes prediction using captioners that have been trained on every other dataset, excluding the dataset in focus. This design aims to test the generalization capabilities of the algorithms when exposed to unfamiliar data, and the possible benefits of using the ensemble in such a scenario.

In Tables VI–IX, a marked decline in the performance of all captioners is observed across every metric and dataset. This is especially true for models trained on the UAV-Captions dataset, which consistently shows the lowest performance metrics, as highlighted by the yellow lines in the tables. Just by looking

CC-a - a large number of trees were planted around the stadium	CC-a - a baseball field is near several green trees
CC-b - many boats are in a port near many buildings and green trees	CC-b - a baseball field is surrounded by many green trees
CC-c - many boats are in a port near many buildings	CC-c - a playground is surrounded by many green trees and buildings
CC-d - many buildings are near a port near many buildings	CC-d - a baseball field is surrounded by some green trees and buildings
MLAT - several boats are in a port near some buildings	MLAT - a baseball field is near several green trees
Blip2 - the harbor of hong kong	Blip2 - aerial view of a baseball field
CapDec - A large body of water with a boat and people on it.	CapDec - A baseball field with a baseball player holding a bat on it.
Naive - several boats are in a port near some buildings	Naive - a baseball field is near several green trees
CLIP-rsicdv2 - many buildings are near a port near many buildings	CLIP-rsicdv2 - a baseball field is surrounded by some green trees and buildings
CLIP-vitlarge14 - several boats are in a port near some buildings	CLIP-vitlarge14 - a baseball field is near several green trees
VaE - many ships are in a port near a beach	VaE - a baseball field is surrounded by many green trees

Fig. 7. Qualitative results on RSICD-Captions, scenario 1.

TABLE IX
UCM-CAPTIONS: GENERALIZATION EVALUATION BLEU1-4 (B1-4), ROUGE (R), METEOR (M), CIDER (C)

	B1	*B2*	*B3*	*B4*	*R*	*M*	*C*
CC-a	0.17	0.07	0.02	0.00	0.17	0.05	0.03
CC-b	0.22	0.11	0.00	0.00	0.19	0.06	0.04
CC-c	0.20	0.09	0.00	0.00	0.18	0.06	0.03
CC-d	0.21	0.09	0.00	0.00	0.19	0.06	0.03
CC-a	0.37	0.17	0.09	0.05	0.27	0.12	0.34
CC-b	0.39	0.19	0.10	0.06	0.28	0.13	0.38
CC-c	0.35	0.18	0.10	0.06	0.26	0.11	0.35
CC-d	0.42	0.20	0.10	0.05	0.28	0.13	0.38
CC-a	0.35	0.17	0.10	0.07	0.29	0.12	0.22
CC-b	0.38	0.19	0.12	0.08	0.28	0.12	0.22
CC-c	0.34	0.20	0.12	0.08	0.29	0.12	0.17
CC-d	0.37	0.22	0.13	0.07	0.32	0.15	0.29
MLAT [29]	0.42	0.23	0.13	0.08	0.31	0.14	0.54
Blip-2 [30]	0.34	0.20	0.12	0.08	0.29	0.13	0.40
CapDec [31]	0.31	0.15	0.06	0.02	0.25	0.11	0.16
Naïve	0.43	0.24	0.14	0.09	0.32	**0.16**	0.47
CLIP	0.40	0.24	0.14	0.08	0.32	0.15	0.49
VaE	**0.47**	**0.32**	**0.22**	**0.16**	**0.38**	**0.16**	**0.62**

Bold entries highlight the best results.

at the metrics, we can deduce that UCM, RSICD, and SYDNEY datasets share certain features or characteristics that make them more closely aligned, while UAV-Captions is substantially different. Among ensemble approaches, the VaE fusion stands out as the most effective method. The unique strength of this approach lies in the capability of the VaE to act as a semantic "translator." The VaE is fine-tuned on the captions of the target

CC-a - there is an asphalted zone	CC-a- there is large road
CC-b - there is soil ground	CC-b- there is white roof
CC-c - there is grass field	CC-c- grass field at upper right is close to asphalt
CC-d - there is large road	CC-d- there is parking lot
CC-a - a piece of ocean is near a yellow beach	CC-a - many planes are parked in an airport
CC-b - it is a piece of yellow desert	CC-b - some boats are in a large port near a road
CC-c - many green trees are in two sides of a curved river	CC-c - several storage tanks are near a piece of green meadow
CC-d - a piece of ocean is near a piece of green ocean	CC-d - many planes are parked in an airport
CC-a - a meadow with some green bushes and white bunkers on it while a highway passed by	CC-a - a small river with dark green waters goes through a residential area
CC-b - a meadow with some green bushes and white bunkers on it	CC-b - a part of ocean with deep green waters
CC-c - there are some white buildings in the industrial area with some roads go through	CC-c - there are some marking lines on the straight runway while some lawns beside
CC-d - a big meadow with some mark lines on it while a highway beside	CC-d - there are some white airplanes parked on the airport with some airport buildings beside
MLAT - a white waves is near a yellow beach	MLAT - many planes are parked in an airport
Blip2 - a man is walking on the beach	Blip2 - a small plane parked on the tarmac
CapDec - a close up of a person wearing a suit and tie	CapDec - A group of planes sitting on top of an airport tarmac.
Naive - a meadow with some green bushes and white bunkers on it while a highway passed by	Naive - many planes are parked in an airport
CLIP-rsicdv2 - a white waves is near a yellow beach	CLIP-rsicdv2 - A group of planes sitting on top of an airport tarmac.
CLIP-vitlarge14 - a man is walking on the beach	CLIP-vitlarge14 - A group of planes sitting on top of an airport tarmac.
VaE - there is a white sand beach with white sands and green waters	VaE - there are two tennis courts on the green ground with a road beside

Fig. 8. Qualitative results on UCM-Captions, scenario 2.

dataset, allowing it to adapt its decoder to the specific vocabulary. This adaption facilitates a sort of semantic distillation through the latent space, thereby "translating" the global meaning of the set of captions into the language style of the target dataset. This improves the relevance of the captions generated, which in turn results in a notable improvement across all performance metrics. The VaE is thus able to focus on the semantic aspects of input captions while overlooking syntactic variations or variations in the choice of words. Unlike other selective ensemble methods, the VaE inherently performs this semantic translation, making it a useful tool for bridging the semantic gap across diverse training datasets. Qualitative results, depicted in Figs. 8 and 9, highlight a difficult scenario for the ensemble, in which most of the captions are unrelated and not coherent with the image. By far the most robust alternative in this case is the CLIP-selection strategy, particularly CLIP-rsicdv2, which provides coherent captions for all the images. Its general, nonspecialized counterpart, CLIP-vitlarge14, is tricked in the first image on the UCM-Captions dataset but provides coherent captions for all the other cases.

CC-a - there is an asphalted zone CC-b - there is an asphalted zone CC-c - there are several trees CC-d - there are several rocks on the asphalt	CC-a- there is road CC-b- large white roof and some grass on the top CC-c- large white roof with shadow on the bottom right CC-d- there is an asphalted zone
CC-a - there are two storage tanks on the ground CC-b - lots of boats docked in lines at the harbor and the water is deep blue CC-c - lots of cars parked in lines in the parking lot CC-d - a big intersection with sky blue roofs	CC-a - it is a small baseball diamond with sand and grass CC-b - it is a small baseball diamond with sand and grass CC-c - a medium residential area with a road goes through this area CC-d - it is a small baseball diamond
CC-a - a meadow with some green bushes and white bunkers on it while a highway passed by CC-b - a meadow with some green bushes and white bunkers on it CC-c - there are some white buildings in the industrial area with some roads go through CC-d - there are some sandlands and orange roofs arranged neatly	CC-a - a residential area with houses arranged neatly and some roads go through this area CC-b - a curved river with dark green waters goes through a residential area CC-c - there are some white marking lines on the runways while some lawns beside CC-d - there are some white buildings in the residential area with some roads go through
Blip2 - aerial view of a harbor with boats CapDec - a number of small boats in a large body of water	Blip2 - a baseball field with a soccer field and a pool CapDec - A baseball player holding a bat on top of a baseball field.
Naive - there are some sandlands and orange roofs arranged neatly CLIP-rsicdv2 - aerial view of a harbor with boats CLIP-vitlarge14 - aerial view of a harbor with boats VaE - there are some cement roads on the desert	Naive - it is a small baseball diamond with sand and grass CLIP-rsicdv2 - a baseball field with a soccer field and a pool CLIP-vitlarge14 - a baseball field with a soccer field and a pool VaE - a bareland with some cars is next to the straight roads

Fig. 9. Qualitative results on RSICD-Captions, scenario 2.

C. Scenario 3: Robustness Evaluation

In this setup, we add noise to the caption set to mimic real-world issues like errors or misspellings. We want to test the resilience of the ensemble to such kind of errors in the input set. To simulate the errors, we decided to operate considering different levels of noise. We represent these levels as a percentage of corrupted words over the total *word count* in the input set of captions. The errors have been simulated using word deletion, word replacing, and character replacing. We decided to test seven noise levels, from 0% to 30% in steps of 5%. Mathematically, the corruption level is represented as

$$\text{Corruption Level} = \frac{\text{Number of Corrupted Words}}{\text{Total Word Count}} \times 100. \quad (2)$$

Errors are simulated using word deletion, word substitution, and character substitution. We evaluate the ensemble's performance across seven discrete noise levels, ranging from 0% to 30%, incremented in steps of 5%.

Results are provided in Tables X– XIII. A key observation is the different performance of selective and generative strategies under varying levels of noise corruption. Specifically, under low noise conditions, selective strategies exhibit superior performance, while under conditions of elevated noise, the VaE generative approach demonstrates a less rapid decrease in performance, highlighting its ability to leverage the global meaning to correct errors. This trend is not observed for the RSICD-Captions dataset. We hypothesize that the bigger size of this dataset increases the variability of captions and scenes, thereby making it more difficult for the VaE to adapt to its distribution. In contrast, smaller datasets present a lower range of variance and complexity, which may render the VaE more effective in adapting to the dataset's inherent distribution. Importantly, our experiments validate the efficacy of ensemble strategies in mitigating the impact of syntactic errors. This suggests that ensemble methods can serve as a robust countermeasure against various

TABLE X
SIDNEY-CAPTIONS: ROBUSTNESS EVALUATION RESULTS EXPRESSED IN TERMS OF BLEU-4

Noise level (%)	*0*	*5*	*10*	*15*	*20*	*25*	*30*
Best captioner	**0.55**	0.50	0.44	0.40	0.34	0.30	0.26
Worst captioner	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Naïve	0.53	**0.53**	**0.49**	**0.45**	0.40	0.35	0.29
CLIP-rsicdv2	0.34	0.33	0.29	0.27	0.26	0.24	0.20
CLIP-vitlarge14	0.33	0.32	0.33	0.30	0.26	0.27	0.20
VaE	0.46	0.46	0.46	**0.45**	**0.46**	**0.44**	**0.41**

Bold entries highlight the best results.

TABLE XI
RSICD-CAPTIONS: ROBUSTNESS EVALUATION RESULTS EXPRESSED IN TERMS OF BLEU-4

Noise level (%)	*0*	*5*	*10*	*15*	*20*	*25*	*30*
Best captioner	**0.32**	0.29	0.25	0.22	0.19	0.17	0.14
Worst captioner	0.02	0.02	0.02	0.02	0.02	0.01	0.01
Naïve	**0.32**	**0.31**	**0.28**	**0.26**	**0.23**	**0.20**	**0.17**
CLIP-rsicdv2	0.26	0.24	0.22	0.20	0.18	0.16	0.14
CLIP-vitlarge14	0.14	0.14	0.14	0.13	0.12	0.11	0.10
VaE	0.26	0.25	0.22	0.19	0.14	0.11	0.08

Bold entries highlight the best results.

TABLE XII
UCM-CAPTIONS: ROBUSTNESS EVALUATION RESULTS EXPRESSED IN TERMS OF BLEU-4

Noise level (%)	*0*	*5*	*10*	*15*	*20*	*25*	*30*
Best captioner	**0.65**	**0.59**	0.52	0.46	0.40	0.36	0.30
Worst captioner	0.04	0.03	0.03	0.02	0.02	0.02	0.02
Naïve	0.60	**0.59**	**0.54**	0.50	0.44	0.38	0.33
CLIP-rsicdv2	0.52	0.48	0.44	0.40	0.37	0.34	0.28
CLIP-vitlarge14	0.34	0.35	0.34	0.33	0.30	0.28	0.24
VaE	0.54	0.52	0.52	**0.51**	**0.47**	**0.45**	**0.42**

Bold entries highlight the best results.

TABLE XIII
UAV-CAPTIONS: ROBUSTNESS EVALUATION RESULTS EXPRESSED IN TERMS OF BLEU-4

Noise level (%)	*0*	*5*	*10*	*15*	*20*	*25*	*30*
Best captioner	0.40	0.37	0.33	0.29	0.24	0.21	0.18
Worst captioner	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Naïve	**0.42**	**0.41**	**0.39**	**0.37**	**0.33**	**0.29**	**0.26**
CLIP-rsicdv2	0.13	0.13	0.13	0.13	0.12	0.12	0.13
CLIP-vitlarge14	0.07	0.08	0.08	0.09	0.09	0.09	0.09
VaE (mean)	0.36	0.35	0.34	0.32	0.31	**0.29**	**0.26**

Bold entries highlight the best results.

TABLE XIV
COMPARATIVE ANALYSIS OF CAPTIONING STRATEGIES

Strategy	Strengths	Drawbacks
Naïve	• Easy to integrate • Good tradeoff between robustness and computational demand	• Unstable with high caption semantic variance • Ignores coherence of the captions during semantic "averaging" in the latent space • Cannot handle syntactic errors
CLIP Selection	• Easy to integrate • Stable with high semantic variance • Selective behavior for image coherence	• Cannot handle syntactic errors • Not tailored for remote sensing images
VaE	• Handles syntactic errors	• Ignores coherence of the captions during semantic "averaging" in the latent space • Nondeterministic output due to generation sampling • Higher computational load

CC-a - this is donnell beach blue sea and white sands	CC-a - it is straight runway with some on it
CC-b - the waves slapping a white sand beach	CC-b - there are straight freeways closen to each other
CC-c - this is g dense forest with green waters and grass	CC-c - there are some mark juarez on the straight runway
CC-d - the waves slapping a white sand beach over and over again	CC-d - there are airplanes in the airport
MLAT - a buyers deemed is near a yellow beach	MLAT - many planes are parked in an airport
Blip2 - a undercover is walking food the beach	Blip2 - aerial viev of a small airport
CapDec - woman in a white dress is standing ip the snow.	CapDec - A group of planes parked in ancestral airport.
Naive - this is donnell beach blue sea and white sands	Naive - there are airplanes in the airport
CLIP-rsicdv2 - the waves slapping a white sand beach over and over again	CLIP-rsicdv2 - A group of planes parked in ancestral airport.
CLIP-vitlarge14 - woman in a white dress is standing ip the snow.	CLIP-vitlarge14 - A group of planes parked in ancestral airport.
VaE - a white sand beach is taxing on one side and a small white sand beach on the other side	VaE - there are some airplanes with black fuselage taxiing on the runway

Fig. 10. Qualitative results on UCM-Captions, scenario 3. Noise level: 20%.

CC-a - a large number of quantitative were planted aroudd the stadium	CC-a - a 130 field is surrounded by some green trees and buildings
CC-b - many boats are in a port near many buildings and khz hrees	CC-b - a field fs stumble by many green trees
CC-c - many boats are in n fort near many buildings	CC-c - o is surrounded by green trees and buildings
CC-d - many buildings are near a port near many budldings	CC-d - a baseball field is surrounded by some green trees and buildings
MLAT - several boats are in a port near soqe buildings	MLAT - a baseball field is near several trees
Blip2 - the harbor of hong impulse	Blip2 - aerial view of a baseball field
CapDec - A restaurant of water with i boat and people on ig.	CapDec - o baseball field with a lo player holding a bat on it.
Naive - several boats are in a port near soqe buildings	Naive - a baseball field is near several trees
CLIP-rsicdv2 - many buildings are near a port near many budldings	CLIP-rsicdv2 - l baseball field is near several breen trees
CLIP-vitlarge14 - the harbor of hong impulse	CLIP-vitlarge14 - aerial view of a baseball field
VaE - many ships in port are orderly surrounded by a port	VaE - a baseball field locates in the meadow with several trees

Fig. 11. Qualitative results on RSICD-Captions, scenario 3. Noise level: 20%.

forms of linguistic noise, thereby enhancing overall resilience. Qualitative results are depicted in Figs. 10 and 11.

V. DISCUSSION

After collecting and analyzing all the results for the three proposed configurations, we summarize our findings and insights on the concept of ensemble in image captioning. As the results demonstrate, an ensemble of captioners can be used to increase the generalization robustness of the output to various situations, scenes, vocabulary, and other factors of variation. Upon the proposed techniques, the main strengths and drawbacks are reported in Table XIV.

The naïve approach, despite its simplicity, proves to be a very strong baseline in this context. The main problem of the naïve approach is the lack of prior filtering on the set of input captions. This problem affects also the VaE framework, which is further affected by the nondeterministic behavior that adds a source of potential errors during caption distillation. The more promising approach in our opinion is the CLIP selection, which provides an effective way of dealing with situations of high variability in the input set. We speculate that the release of more targeted CLIP models for the remote sensing scenario can benefit the selection ability, and thus, improve the robustness of

TABLE XV
SINGLE MODEL CAPTION GENERATION TIME (SECONDS PER IMAGE)

Model Type	Time (s)
CC-a	0.025
CC-b	0.030
CC-c	0.126
CC-d	0.140
MLAT	2.660
Blip-2	0.654
CapDec	0.160
Total generation time	3.795

TABLE XVI
ENSEMBLE FUSTION STRATEGY ADDITIONAL TIME (SECONDS PER IMAGE), TOTAL TIME, AND % OVER TOTAL TIME

Fusion strategy	Time (s)	Total time (s)	% of Total time
Naïve	0.010	3.805	0.26%
CLIP coherence	0.038	3.833	0.99%
VAE	0.250	4.045	6.18%

the CLIP selection ensemble strategy. Furthermore, we analyzed the additional computation time required to run our ensemble. This mainly depends on the number of models included in the ensemble, with little additional computation time required by the selection/fusion methods. Table XV reports the time used by each captioner to produce a caption for an image, and Table XVI reports the additional computational time for the fusion strategy, along with the total time. As can be seen, running the ensemble requires the generation of a caption by each model plus the added time for the fusion strategy. This leads to a severe increase in computational time, with the slowest method being VaE with 4.045 seconds per image.

VI. CONCLUSION

In this study, we have systematically investigated the generation and application of an ensemble approach to increase the robustness of image captioning. Specifically, the strategies employed were a posteriori, acting after individual captions have been generated. Two strategies, naïve and CLIP-selection, focus on choosing the most coherent caption from a generated set, which we refer to as selective captioning. The other strategy, based on the VaE framework synthesizes a new caption based on the entire set of generated captions, which we refer to as generative captioning. We conducted a comprehensive analysis on three well-established remote sensing image captioning datasets in addition to a novel dataset, termed UAV-Captions. Three scenarios have been designed to test different aspects of the captioning process, which allowed us to expose and discuss the strengths and weaknesses of each implemented method. Our findings demonstrate that our ensemble-based approaches offer a scalable and robust pipeline for integrating various captioning algorithms. This leads to more reliable and contextually accurate captions. More specifically, our results suggest that the CLIP coherence selection is less sensible to noisy and unrelated captions, and thus, more suitable in situations in which there is a high variation in the generated caption set. The approach based on the VaE has shown to be robust to noise, but at low noise levels, it is outperformed by the selective strategies. In addition, we show that the use of multiple captioners incurs a significant computational overhead with respect to single-model alternatives, making the use of the ensemble suitable only when time is not a constraint. Avenues for future research in this area include the integration of CLIP models specifically tailored for remote sensing imagery [39], as well as the implementation of automated filtering mechanisms to prune less coherent caption candidates before ensemble application. The integration of tailored CLIP models can reduce the leaking of irrelevant captions as happened in Fig. 10 for the CLIP-vitlarge14. This model is trained on natural images, and we can see that from the colors in the image, the model is tricked into describing the image as a "woman in white dress." The more the CLIP model is tailored to the RS scenario, the less the selection of irrelevant captions. On the other hand, a mechanism to filter the captions before naïve and VaE solution can benefit both strategies, by removing a portion of irrelevant captions that could contaminate the selection and distillation with unrelated information. In summary, the ensemble strategies examined in this article hold promise for significantly enhancing the reliability and contextual relevance of image captions in remote sensing applications.

REFERENCES

[1] J. Quan, C. Wu, H. Wang, and Z. Wang, "Structural alignment based zero-shot classification for remote sensing scenes," in *Proc. IEEE Int. Conf. Electron. Commun. Eng.*, 2018, pp. 17–21.
[2] C. Wang, G. Peng, and B. De Baets, "A distance-constrained semantic autoencoder for zero-shot remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12545–12556, Dec. 2021.
[3] Y. Chen, C. Wei, D. Wang, C. Ji, and B. Li, "Semi-supervised contrastive learning for few-shot segmentation of remote sensing images," *Remote Sens.*, vol. 14, no. 17, 2022, Art. no. 4254.
[4] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.
[5] Z. Dong, M. Wang, Y. Wang, Y. Zhu, and Z. Zhang, "Object detection in high resolution remote sensing imagery based on convolutional neural networks with suitable object scale features," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2104–2114, Mar. 2020.
[6] G. Hoxha, F. Melgani, and B. Demir, "Toward remote sensing image retrieval under a deep image captioning perspective," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4462–4475, Aug. 2020.
[7] Z. Yuan et al., "Remote sensing cross-modal text-image retrieval based on global and local information," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5620616.
[8] R. Zhao and Z. Shi, "Text-to-remote-sensing-image generation with structured generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Mar. 2021, Art. no. 8010005.
[9] M. B. Bejiga, F. Melgani, and A. Vascotto, "Retro-remote sensing: Generating images from ancient texts," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 3, pp. 950–960, Mar. 2019.
[10] 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, May 2021, Art. no. 5606514.
[11] R. Ricci, Y. Bazi, and F. Melgani, "Machine-to-machine visual dialoguing with ChatGPT for enriched textual image description," *Remote Sens.*, vol. 16, no. 3, 2024, Art. no. 441.
[12] S. He, W. Liao, H. R. Tavakoli, M. Y. Yang, B. Rosenhahn, and N. Pugeault, "Image captioning through image transformer," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 153–169.
[13] X. Yang et al., "Fashion captioning: Towards generating accurate descriptions with semantic rewards," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 1–17.

[14] Q. Xia et al., "XGPT: Cross-modal generative pre-training for image captioning," in *Proc. Natural Lang. Process. Chin. Comput.*, 2020, pp. 786–797.
[15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3156–3164.
[16] 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.
[17] 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.
[18] X. Chen et al., "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*.
[19] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, 2014.
[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, May 7–9, 2015.
[21] 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.
[22] Q. Yang, Z. Ni, and P. Ren, "Meta captioning: A meta learning based remote sensing image captioning framework," *ISPRS J. Photogrammetry Remote Sens.*, vol. 186, pp. 190–200, 2022.
[23] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process Syst.*, vol. 30, 2017.
[24] J. Wang, Z. Chen, A. Ma, and Y. Zhong, "CapFormer: Pure transformer for remote sensing image caption," in *Proc. IGARSS IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 7996–7999.
[25] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.
[26] G. Briem, J. Benediktsson, and J. Sveinsson, "Multiple classifiers applied to multisource remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2291–2299, Oct. 2002.
[27] F. Melgani and Y. Bazi, "Markovian fusion approach to robust unsupervised change detection in remotely sensed imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 4, pp. 457–461, Oct. 2006.
[28] H. Katpally and A. Bansal, "Ensemble learning on deep neural networks for image caption generation," in *Proc. IEEE 14th Int. Conf. Semantic Comput.*, 2020, pp. 61–68.
[29] C. Liu, R. Zhao, and Z. Shi, "Remote-sensing image captioning based on multilayer aggregated transformer," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Feb. 2022, Art. no. 6506605.
[30] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19730–19742.
[31] D. Nukrai, R. Mokady, and A. Globerson, "Text-only training for image captioning using noise-injected clip," *Findings Assoc. Comput. Linguistics*, Abu Dhabi, United Arab Emirates, pp. 4055–4063, Dec. 7–11, 2022, doi: 10.18653/V1/2022.FINDINGS-EMNLP.299.
[32] H. W. Chung et al., "Scaling instruction-finetuned language models," *J. Mach. Learn. Res.*, vol. 25, no. 70, pp. 1–53, 2024.
[33] B. Zhao, "A systematic survey of remote sensing image captioning," *IEEE Access*, vol. 9, pp. 154086–154111, 2021.
[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.
[35] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 5776–5788, 2020.
[36] Huggingface, "Sentence-transformers/all-minilm-l6-v2," [Online]. Available: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[37] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, pp. 8748–8763, Jul. 18–24, 2021.
[38] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Representations*, May 3–7, 2021.
[39] F. Liu et al., "Remoteclip: A vision language foundation model for remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, Apr. 2024.
[40] Huggingface, "openai/clip-vit-large-patch14." [Online]. Available: https://huggingface.co/openai/clip-vit-large-patch14
[41] Huggingface, "flax-community/clip-rsicd-v2." [Online]. Available: https://huggingface.co/flax-community/clip-rsicd-v2
[42] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Józefowicz, and S. Bengio, "Generating sentences from a continuous space," in *Proc. 20th Conf. Computat. Natural Lang. Learn.*, Aug. 2016, pp. 10–21, doi: 10.18653/v1/K16-1002.
[43] C. Li et al., "Optimus: Organizing sentences via pre-trained modeling of a latent space," 2020, *arXiv:2004.04092*.
[44] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
[45] G. Hoxha and F. Melgani, "A novel SVM-based decoder for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2021, Art. no. 5404514.
[46] 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*, 2002, pp. 311–318.
[47] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2004, pp. 74–81.
[48] A. Lavie and A. Agarwal, "Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Stat. Mach. Transl.*, 2007, pp. 228–231.
[49] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.

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 in the field of large language models and large multimodal models for remote sensing applications, focusing on the intersection of natural language processing and computer vision applied to remote sensing.

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.

Farid Melgani (Fellow, IEEE) received the State Engineer degree in electronics from the University of Batna, 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 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, and is also the Head of the Signal Processing and Recognition Laboratory and the Dean of Undergrad and Grad Studies. He is coauthor of more than 270 scientific publications. His research interests include the areas of remote sensing, signal/image processing, and artificial intelligence.

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

José Marcato Junior (Member, IEEE) received the Ph.D. degree in cartographic science from Sao Paulo State University, São Paulo, Brazil, in 2014.

He is currently a Professor with the Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul, Campo Grande, Brazil. He has authored and coauthored more than 100 research papers in refereed journals and more than 70 in conferences, including *ISPRS Journal of Photogrammetry and Remote Sensing*, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, and IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING. His research interests include the integration of Remote Sensing and Photogrammetry with Deep Learning for object detection, classification, and segmentation.

Wesley Nunes Gonçalves received the B.Sc. degree in computer engineering from Dom Bosco Catholic University, Campo Grande, Brazil, in 2007, and the M.Sc. degree in computer science and the Ph.D. degree in computational physics from the University of São Paulo, São Paulo, Brazil, in 2010 and 2013, respectively.

He is currently a Professor with the Federal University of Mato Grosso do Sul, Campo Grande. He has authored several refereed papers. His research interests include object detection, image segmentation, and the application of computational methods in remote sensing and precision agriculture.