WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal Homepage: www.ijrpr.com ISSN 2582-7421

Vision-Language Models for Satellite Image Understanding and Natural-Language Reporting: A Survey and a Lightweight LoRA-Based Captioning Pipeline

A. Ratna Rishika[1] and Dr. S. Akila Agnes[2]

[1]B. Tech Student (Reg. No. 24341A0507), Department of Computer Science and Engineering, GMRIT Deemed to be University, Rajam, Andhra Pradesh, India – 532127

[2]Associate Professor, Department of Computer Science and Engineering, GMRIT Deemed to be University, Rajam, Andhra Pradesh, India – 532127

ABSTRACT—

The rapid growth of high-resolution satellite data has made manual interpretation slow and dependent on expert analysts. Vision-Language Models (VLMs) address this problem by combining computer vision with natural language processing, so that an image can be described, questioned and reported on in plain language. This paper surveys twenty recent publications on vision-language learning for remote sensing, grouped into foundation models, large instruction-tuned multimodal models, image–text datasets, cross-modal alignment and task-specific studies. The surveyed work is compared by objective, limitation, advantage, evaluation metric and remaining research gap, and five representative methodologies (AddressCLIP, LHRS-Bot, VHM, a generative change-detection framework and SARCLIP) are analysed in detail. On this basis, a lightweight pipeline is proposed in which a pre-trained BLIP captioning model is adapted to remote-sensing imagery using Low-Rank Adaptation (LoRA) on the RSICD dataset. On 200 test images the BLEU score improved from 3.92 for the un-tuned baseline to 10.23 for the fine-tuned model, and the captions changed from a generic template to descriptions that name objects and layout. Applications, open challenges (scarce captions, domain gap, model size and hallucination) and future directions are also discussed.

***Keywords*—Vision-Language Models, Satellite Image Understanding, Remote Sensing, Image Captioning, Multimodal Artificial Intelligence, Earth Observation, LoRA**

I. INTRODUCTION

Satellite imagery supports environmental monitoring, disaster management, urban planning, precision agriculture and climate studies. The volume of highresolution data now grows faster than it can be interpreted manually. Conventional remote-sensing pipelines concentrate on scene classification or object detection, and their outputs (class labels and bounding boxes) still have to be interpreted by a trained analyst before they are useful to a decision-maker.

Vision-Language Models (VLMs) learn from large collections of image–text pairs and support image captioning, visual question answering (VQA), scene understanding and cross-modal retrieval [3], [4]. By linking visual and textual information they make remote-sensing analysis accessible to non-experts and enable automatic natural-language reporting [2]. Since the introduction of CLIP [4], a large body of work has adapted this idea to overhead imagery, producing remote-sensing-specific foundation models, instruction-tuned multimodal language models and large image–text datasets.

The contributions of this paper are:

1. A structured survey of twenty vision-language papers for remote sensing, compared by objective, limitation, advantage, metric and research gap;
2. A detailed analysis of five representative methodologies, including the base paper addressclip [1];
3. Identification of five research gaps: data scarcity, domain adaptation, hallucination, model size and the lack of complete report generation; and
4. A lightweight pipeline that fine-tunes blip with lora on rsicd and reports the improvement over the un-tuned baseline.

Section II reviews the literature, Section III describes the methodology, Section IV analyses representative existing methods, Section V presents the implementation and results, Section VI discusses applications and challenges, and Section VII concludes.

II. LITERATURE SURVEY The twenty reviewed papers are grouped by their main contribution.

A. Foundations and surveys.

Radford et al. [4] introduced CLIP, which learns visual representations from natural-language supervision by aligning images and text in a shared embedding space; this enables zeroshot recognition and transfer learning and underlies almost all later remote-sensing VLMs. Li et al. [3] review vision-language models in remote sensing, covering captioning, VQA, open challenges and future trends. The base paper of this work, AddressCLIP [1], extends CLIP to image address localization using image–address and image–caption contrastive learning together with an image–geography matching loss.

B. Large multimodal and instruction-tuned models.

LHRS-Bot [5] uses volunteered geographic information (VGI) to improve vision–language alignment. VHM [6] adds honest instruction tuning to reduce hallucination. RSGPT [10] provides a large remote-sensing VLM with a benchmark, and SkyEyeGPT [11] unifies captioning, VQA, grounding and report generation in one instruction-tuned framework. GeoChat [18] supports image-level and region-level dialogue and object grounding, and RS-LLaVA [21] performs joint captioning and question answering with the RS-Instructions dataset.

C. Datasets and image–text corpora.

RS5M with GeoRSCLIP [19] and SkyScript [20] provide largescale remote-sensing image–text data for CLIP-style models. RSTeller [15] uses language-model-written captions with rich linguistic semantics, and LuoJiaHOG [16] contributes hierarchy-oriented, geo-aware captions for retrieval. Cha et al. [9] remove the need for human annotation by generating image–text datasets automatically.

D. Alignment and modality-specific models.

Zavras et al. [12] study the modality gap between satellite images and text to improve cross-modal alignment, while SARCLIP [8] is the first vision-language foundation model for Synthetic Aperture Radar (SAR) imagery.

E. Task-specific studies.

Choudhury et al. [13] improve visual grounding with adaptive modality guidance, Zhang et al. [17] use frequency-domain transfer learning for remote-sensing VQA, Chen et al. [14] propose a multimodal retrieval-augmented generation framework for lithological recognition, and Zan et al. [7] use open-vocabulary generative VLMs to build a large change-detection dataset.

TABLE I. Summary of representative surveyed work

Ref.	Work	Main limitation	Research gap
[1]	AddressCLIP	Street-level, not satellite imagery	Extend to satellite reporting
[4]	CLIP	General-domain model	Domain adaptation
[5]	LHRS-Bot	Complex, diverse scenes	Region-level reasoning
[6]	VHM	Large model size	Efficient deployment
[8]	SARCLIP	SAR images only	Optical–SAR fusion
[10]	RSGPT	High compute cost	Real-time use
[11]	SkyEyeGPT	Complex instruction tuning	Lightweight tuning
[18]	GeoChat	Large model	Efficient region reasoning
[19]	RS5M / GeoRSCLIP	Large data and compute	Higher-quality captions
[20]	SkyScript	Noisy automatic captions	Cleaner captions
[21]	RS-LLaVA	Needs instruction data	Lighter models

TABLE II. Datasets and benchmarks introduced in the surveyed papers

Dataset	**Ref.**	**Content and purpose**
IAL datasets	[1]	Street-level images with address text; address localization
LHRS-Align	[5]	~1.15M image–text pairs from VGI; alignment
RSICap / RSIEval	[10]	2,585 human-described images; fine-tuning and benchmark
SkyEye-968k	[11]	Multi-task instruction data; instruction tuning
RSTeller	[15]	LLM-written captions; pre-training
LuoJiaHOG	[16]	Hierarchical geo-aware captions; retrieval benchmark
GeoChat set	[18]	~306K image–instruction pairs; instruction tuning
RS5M	[19]	~5M filtered or auto-captioned pairs; pretraining
SkyScript	[20]	~2.6M pairs linking Earth Engine images with OSM tags
RS-Instructions	[21]	Instructions from captioning and VQA data

Captions are obtained in three ways: by converting existing annotations into text, such as detection labels for SAR images [8] or OpenStreetMap tags for optical images [5], [20]; by using a captioning model or a large language model [15], [19]; and, in only a few cases, by human annotation [10]. This difference explains the trade-off between scale and quality discussed below.

TABLE III. Evaluation metrics typically used for each task

Task	**Typical metrics**	**What is measured**
Image captioning	BLEU, METEOR, ROUGE, CIDEr	Overlap with reference captions
VQA	Accuracy	Share of correct answers
Zero-shot classification	Top-1 accuracy	Correct class ranked first
Image–text retrieval	Recall@K, precision	Correct match in top K
Visual grounding	Accuracy at IoU threshold	Region overlap
Change detection	Precision, recall, F1, mIoU	Pixel-level agreement

F. Critical analysis. The literature shows a clear progression: CLIP [4] established natural-language supervision; remote-sensing datasets [15], [16], [19], [20] then grew in size and richness; and recent models [5], [6], [10], [11], [18], [21] connect vision encoders to large language models so that the output is free text. Five observations follow. First, data is the dominant bottleneck, and the choice between a few high-quality human captions [10] and many automatic captions [9], [19], [20] is a trade-off between fluency and scale. Second, general models need domain adaptation, usually through parameter-efficient methods such as LoRA. Third, hallucination is rarely measured; only a few works [6], [14] address it directly. Fourth, large models limit deployment on modest hardware. Fifth, few works deliver a complete report combining scene type, objects, layout and confidence; this is addressed only in part [11]. These gaps motivate a practical, lightweight pipeline that adapts a pre-trained VLM to satellite imagery with limited resources and presents the result as a readable report.

III. METHODOLOGY AND SYSTEM ARCHITECTURE

The proposed methodology converts a satellite image into a natural-language description in six stages, shown in Fig. 1.

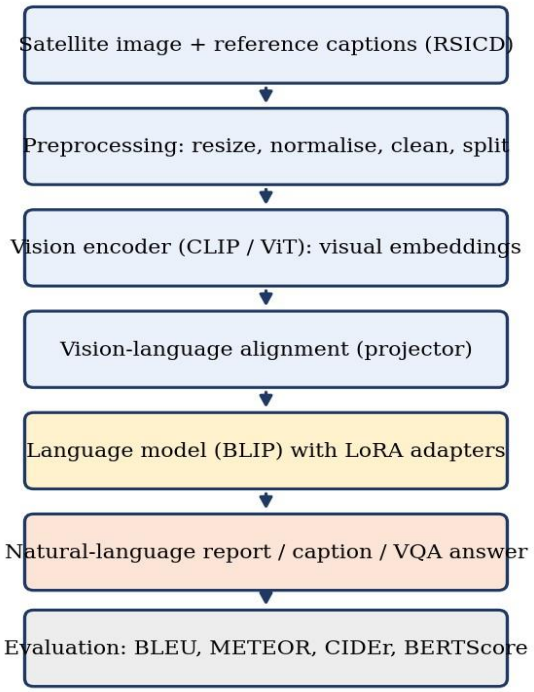


Fig. 1. Architecture of the proposed satellite image-to-report pipeline

A. *Image acquisition and preprocessing.*

Images and reference captions are taken from public remote-sensing datasets such as RSICD and UCM-Captions, which cover buildings, roads, farmland, forests, rivers and urban regions. Images are resized to the model input size, pixel values are normalised, corrupted or duplicate images are removed, and the data are split into training, validation and test sets.

B. *Visual feature extraction.*

A pre-trained vision encoder (CLIP, ViT or a remote-sensing-specific encoder) converts each image into embeddings that represent objects and spatial patterns such as roads, buildings, vegetation and water.

C. *Vision-language alignment.*

The visual features are aligned with language representations so that the system learns the chain: satellite image → visual features → semantic understanding → textual representation.

D. *Report generation.*

A language model generates a description of the major objects, land-cover type, spatial relationships and overall scene, for example: "The satellite image shows an urban area containing several buildings connected by road networks, with vegetation distributed around the developed region."

E. *Training and fine-tuning.*

The model is fine-tuned on image–text pairs to minimise the difference between generated and reference captions. Instruction tuning can further teach the model to handle different questions and report requests.

F. *Evaluation.*

Generated text is scored with BLEU, ROUGE, METEOR, CIDEr and BERTScore, and human evaluation can assess factual correctness, relevance, completeness and readability.

IV. ANALYSIS OF REPRESENTATIVE EXISTING METHODS

A. *AddressCLIP (base paper) [1].*

Built on CLIP encoders, AddressCLIP combines image–address contrastive learning, image–caption contrastive learning and image–geography matching, and predicts the candidate address with the highest similarity to the image. Three datasets were created: Pitts-IAL (234K images), SF-IAL-Base (184K) and SF-IAL-Large (1.96M). It reaches more than 80% Top-1 accuracy on all three, 85.92% on SF-IAL-Large, and improves SSA1 over

CLIP-based transfer methods by about 7.41%, 4.86% and 6.29% respectively. It shows that textual and geographical supervision make an image model location-aware, but it targets street-level images and not satellite scene reporting.

B. *LHRS-Bot [5].*

LHRS-Bot connects a multi-level vision perceiver to a LLaMA2-7B language model, allocates more queries to lower-level visual features, and uses a three-stage curriculum from alignment to instruction following. It is trained with LHRS-Align (about 1.15M image–text pairs from VGI/OpenStreetMap), LHRS-Instruct and the LHRSBench benchmark. In an ablation, LHRS-Align, although smaller than RS5M and SkyScript, gave the best result on all four downstream metrics, showing that data quality and domain relevance can matter more than size.

C. *VHM [6].*

VHM is trained on rich image–text data and on *honest* instruction tuning, in which it learns to state that an object is not visible instead of inventing an answer. This reduces hallucination and improves reliability, at the cost of a large model.

D. *Zan et al. [7].*

CLIP, a fine-tuned Stable Diffusion model and SD-Inpainting-RS generate synthetic image pairs with exact change labels. The resulting WHUGCD dataset has six land-cover categories and 25 change directions, and models pre-trained on it generalise well to real change-detection datasets.

E. *SARCLIP [8].*

A Detection-to-Caption (D2C) algorithm converts SAR object-detection annotations into captions, forming the MMSAR dataset, and an AnyReader module handles high-resolution SAR images with less information loss. SARCLIP achieves 7.07% higher average image-to-text recall and 7.8% higher zero-shot classification accuracy than CLIP.

V. PROPOSED IMPLEMENTATION AND RESULTS

A. *Setup.*

The pipeline fine-tunes the pre-trained BLIP captioning model on RSICD satellite images with human-written captions, using LoRA with rank r = 16, scaling α = 32 and adapters on the query and value projections. CLIP is used for classification and retrieval, and a Gradio demo exposes captioning and VQA. Training runs on a free Google Colab NVIDIA T4 GPU (16 GB) for three epochs, in roughly 30–60 minutes; inference runs on a laptop CPU with 8 GB RAM. The software stack is Python, PyTorch, Hugging Face Transformers, PEFT, Datasets, Gradio and SacreBLEU.

B. *Baseline.*

Before fine-tuning, the pre-trained model was evaluated on 200 test images and obtained **BLEU = 3.92**. It recognises the scene correctly but always uses the template “an aerial view of …” and omits details that human annotators mention, such as grass, bare land, planes and size.

C. *Fine-tuned model.*

After LoRA fine-tuning on RSICD, the BLEU score on the same 200 test images rose to **10.23** (Table IV) and Fig. 2.

TABLE IV. Baseline versus fine-tuned model (200 test images)

Model	**BLEU**	**Example caption**
Baseline (BLIP, no fine-tuning)	3.92	an aerial view of an airport
Fine-tuned (LoRA on RSICD)	10.23	many planes are parked in an airport near some buildings and some bareland

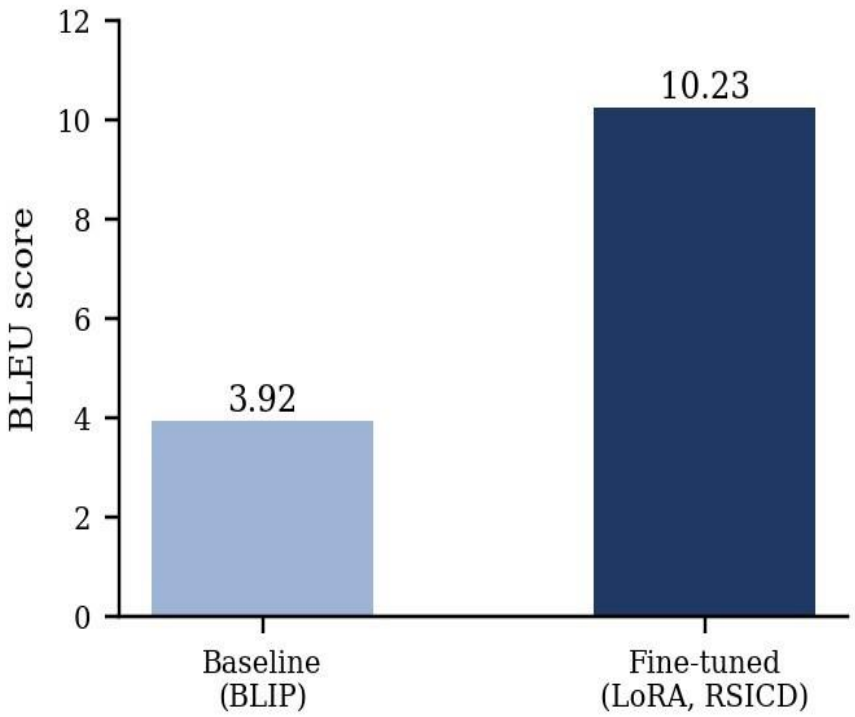


Fig. 2. BLEU score before and after LoRA fine-tuning (200 test images)

The fine-tuned captions mention objects and layout, moving the output closer to human wording. Two case studies illustrate the reporting stage: for a forest image the system predicted *Forest* with 91.51% confidence and reported "a forested region dominated by dense vegetation", and for farmland it predicted *Annual Crop* and described an agricultural region likely used for crop production.

D. *Discussion.*

The BLEU gain shows that parameterefficient domain adaptation is effective even on a free GPU, in line with GeoRSCLIP [19], RS-LLaVA [21] and GeoChat [18]. BLEU remains modest in absolute terms for two reasons: RSICD captions are short and generic, which limits the style of generated reports, a trade-off also noted for automatically produced corpora [15], [20]; and n-gram metrics penalise correct but differently worded captions. Future evaluation should add METEOR, CIDEr, BERTScore and human checks of factual correctness. The system also describes whole images only and does not use region-level grounding [13], [18].

VI. APPLICATIONS, CHALLENGES AND FUTURE DIRECTIONS

Applications. Disaster assessment (describing damaged areas after floods, fires or earthquakes), environmental monitoring and climate analysis, urban planning and land-use monitoring, agriculture and crop monitoring, and geological analysis through retrieval-augmented recognition [14].

Challenges. (1) Limited high-quality annotated image– text datasets and errors in automatically generated captions [9]. (2) Domain gaps between sensors and modalities such as optical and SAR [8]. (3)

Computational cost and model size that hinder real-time use [6], [10], [12]. (4) Hallucination, which is serious for reports because non-existent objects can mislead decision-makers [6].

Future directions. Larger and more diverse remotesensing datasets with better automatic captioning; improved multimodal alignment; lightweight instruction tuning for real-time deployment; explainable, factually grounded report generation; and extensions of the present pipeline to RSVQA finetuning, BLIP-2 and multispectral data.

VII. CONCLUSION

Vision-Language Models provide an effective route from satellite pixels to human-readable reports. This paper surveyed twenty recent works, analysed five representative methods and showed that foundation models, instruction-tuned multimodal models and large image–text datasets together form the current state of the art, while data cost, domain gap, model size and hallucination remain open. A lightweight pipeline that fine-tunes BLIP with LoRA on RSICD improved BLEU from 3.92 to 10.23 on 200 test images and produced captions that describe objects and layout rather than a generic template. Reliable deployment will still require suitable datasets, domain-specific finetuning, richer evaluation metrics and human verification of critical outputs.

ACKNOWLEDGMENT

The authors thank Dr. K. Lakshmanarao (Professor and Head), Mr. Y. Nagapramodkumar (Course Coordinator) and the Department of Computer Science and Engineering, GMRIT Deemed to be University, Rajam, for their guidance and support.

REFERENCES

[1] S. Xu, C. Zhang, L. Fan, G. Meng, S. Xiang, and J. Ye, "AddressCLIP: Empowering vision-language models for citywide image address localization," in *Proc. European Conf. Computer Vision (ECCV)*, 2024, pp. 76–92.

[2] X. Weng, C. Pang, and G.-S. Xia, "Vision-language modeling meets remote sensing: Models, datasets and perspectives," *IEEE Geoscience and Remote Sensing Magazine*, 2025.

[3] X. Li, C. Wen, Y. Hu, Z. Yuan, and X. X. Zhu, "Vision-language models in remote sensing: Current progress and future trends," *IEEE Geoscience and Remote Sensing Magazine*, vol. 12, no. 2, pp. 32–66, 2024.

[4] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Machine Learning (ICML)*, PMLR vol. 139, 2021, pp. 8748–8763.

[5] D. Muhtar, Z. Li, F. Gu, X. Zhang, and P. Xiao, "LHRS-Bot: Empowering remote sensing with VGI-enhanced large multimodal language model," in *Computer Vision – ECCV 2024*, Springer, pp. 440–457.

[6] C. Pang et al., "VHM: Versatile and honest vision language model for remote sensing image analysis," *Proc. AAAI Conf. Artificial Intelligence*, vol. 39, no. 6, pp. 6381–6388, 2025.

[7] Y. Zan, S. Ji, S. Chao, and M. Luo, "Open-vocabulary generative vision-language models for creating a large-scale remote sensing change detection dataset," *ISPRS J. Photogrammetry and Remote Sensing*, vol. 225, 2025.

[8] "SARCLIP: The first vision-language foundation model for SAR image," *IEEE Trans. Geoscience and Remote Sensing*, 2025.

[9] K. Cha, D. Yu, J. Seo, H. Choi, and T. Jeon, "Pushing the limits of vision-language models in remote sensing without human annotations," *Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLVIII-G2025, pp. 249–254, 2025.

[10] Y. Hu, J. Yuan, C. Wen, X. Lu, et al., "RSGPT: A remote sensing vision language model and benchmark," *ISPRS J. Photogrammetry and Remote Sensing*, vol. 224, pp. 272–286, 2025.

[11] Y. Zhan, Z. Xiong, and Y. Yuan, "SkyEyeGPT: Unifying remote sensing vision-language tasks via instruction tuning with large language model," *ISPRS J. Photogrammetry and Remote Sensing*, vol. 221, 2025.

[12] A. Zavras, D. Michail, B. Demir, and I. Papoutsis, "Mind the modality gap: Towards a remote sensing vision-language model via cross-modal alignment," *ISPRS J. Photogrammetry and Remote Sensing*, vol. 228, pp. 270–287, 2025.

[13] S. Choudhury, P. Kurkure, and B. Banerjee, "Improving visual grounding in remote sensing images with adaptive modality guidance," *ISPRS J. Photogrammetry and Remote Sensing*, vol. 224, pp. 42–58, 2025.

[14] X. Chen et al., "A vision-language foundation model-based multimodal retrieval-augmented generation framework for remote sensing lithological recognition," *ISPRS J. Photogrammetry and Remote Sensing*, 2025.

[15] J. Ge, X. Zhang, Y. Zheng, K. Guo, and J. Liang, "RSTeller: Scaling up visual language modeling in remote sensing with rich linguistic semantics," *ISPRS J. Photogrammetry and Remote Sensing*, 2025.

[16] Y. Zhao, M. Zhang, B. Yang, et al., "LuoJiaHOG: A hierarchyoriented geo-aware image caption dataset for remote sensing image-text retrieval," *ISPRS J. Photogrammetry and Remote Sensing*, 2025.

[17] Z. Zhang et al., "Frequency domain transfer learning for remote sensing visual question answering," *Expert Systems with Applications*, vol. 291, p. 128395, 2025.

[18] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "GeoChat: Grounded large vision-language model for remote sensing," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 27831–27840.

[19] Z. Zhang, T. Zhao, Y. Guo, and J. Yin, "RS5M and GeoRSCLIP: A large-scale vision-language dataset and a large vision-language model for remote sensing," *IEEE Trans. Geoscience and Remote Sensing*, vol. 62, pp. 1–23, 2024.

[20] Z. Wang, R. Prabha, T. Huang, J. Wu, and R. Rajagopal, "SkyScript: A large and semantically diverse vision-language dataset for remote sensing," *Proc. AAAI Conf. Artificial Intelligence*, vol. 38, no. 6, pp. 5805–5813, 2024.

[21] F. Liu et al., "RS-LLaVA: A large vision-language model for joint captioning and question answering in remote sensing imagery," *Remote Sensing*, vol. 16, no. 9, p. 1477, 2024.