Exposé - Master's Thesis

Rephrase vision-language modeling in remote sensing

Mark Spitzner
spitznem@informatik.hu-berlin.de
Humboldt University Berlin, Germany

Supervisors:
Prof. Dr. Pedram Ghamisi
p.ghamisi@gmail.com
Helmholtz-Zentrum Dresden-Rossendorf, Germany
Institute of Advanced Research in Artificial Intelligence, Austria

Prof. Dr. Ulf Leser
leser@informatik.hu-berlin.de
Humboldt University Berlin, Germany

January 29, 2025

Contents

1 Introduction **1**

2 Related work **1**

3 Goal and Approach **2**
3.1 Goal and Research Questions 3
3.2 Approach 3

4 Datasets and Evaluation **4**
4.1 Datasets 4
4.2 Evaluation 5

References

Appendices

A Datasets

B Metrics

1 Introduction

Recent advancements in deep learning, particularly within the fields of natural language processing (NLP) and computer vision (CV), coupled with the emergence of multi-modal approaches, have initiated significant research efforts in the area of vision-language tasks (VL tasks). These tasks, e.g. image captioning (IC)[1] (given an image, describe the content in natural language), change captioning (CC) [2] (given images of the same region at different times, describe the changes if any occurred) and visual question answering (VQA) [3] (given an image and a question in natural language, answer the question in natural language in the context of the image content), are significant due to their capability to express a higher degree of domain comprehension. This is evident from the fact that even relatively simple tasks, such as image classification can be reformulated as VQA and solved by a vision-language model (VLM).

In addition, vision-language tasks allow the interaction of non-experts with domain specific data through an intuitive NLP-interface. Especially, in the domain of remote sensing (RS) this is of high importance as there are often high risks involved and personnel might not be trained in the evaluation of RS data.

However, there are various different VL tasks that try to solve different objectives. As a result, the users needed to change their model based on the task they are trying to solve. In order to provide models that excell independent of the task, multiple approaches have aimed to consolidate multiple VL tasks in order to use a single model architecture [4].

In the associated master's thesis we aim to unify different VL tasks. We will reformulate different VL tasks e.g. IC, scene classification (SC) or visual grounding (VG) to fit the concept of VQA. This paradigm is then used to train a model using a Query-Former[5]-like approach to prove its general feasibility. As a second goal we will refine this model architecture using smaller models like DistilGPT2 [6] or GPT-NeoX[7] as language models and MobileViT[8] or Efficient-Former [9] as visual encoder to evaluate the performance of parameter-efficient approaches to allow for a reduced carbon-dioxide footprints.

2 Related work

With RSGPT, Hu et al.[10] introduced a vision-language framework specifically designed for remote sensing, leveraging a domain-adapted vision backbone and a text generation module fine-tuned on a hand-crafted dataset consisting of high-quality annotated image-text pairs. The model achieves state-of-the-art performance in IC and VQA, providing robust capabilities for automated reporting and semantic understanding in geospatial contexts.

Kuckreja et al. proposed GeoChat[11]. Their approach extends large-scale VLM to support region-specific dialogue by incorporating spatial reasoning through a region proposal network. This approach enables precise, localized

queries and contextual descriptions, with applications in disaster response, urban planning and land-use monitoring.

Addressing the limitations of general VLM in handling domain-specific challenges, RemoteCLIP by Liu et al.[12] adapts the CLIP [13] architecture for remote sensing. RemoteCLIP utilizes contrastive pretraining on remote sensing datasets, achieving superior zero-shot performance on image-text alignment tasks. Its robustness to multi-sensor data and ability to generalize across varied resolutions provides a versatile foundation for remote sensing applications.

Li et al. integrated temporal information into VLM in UniRS[14], which unifies multi-temporal remote sensing tasks, including change detection and temporal scene classification, within a single framework. By leveraging temporal embeddings alongside vision-language pretraining, UniRS demonstrates the capacity to model spatiotemporal dependencies, expanding the applicability of VLM to dynamic geospatial scenarios.

To tackle the heterogeneity in spatial resolution inherent in remote sensing, Liu et al. proposed RSUniVLM[15], a mixture-of-experts architecture that dynamically selects specialized modules for coarse- and fine-grained tasks. This approach achieves state-of-the-art results across multiple benchmarks, including object detection and semantic segmentation, while maintaining computational efficiency.

Finally, Zhan et al.[16] unified diverse remote sensing tasks using SkyEyeGPT an instruction-tuned framework, enabling interaction via natural language prompts. By aligning task-specific objectives through instruction tuning, SkyEyeGPT demonstrates unprecedented versatility, handling tasks such as captioning, retrieval and semantic segmentation within a single interface.

These works collectively highlight the significant strides made in adapting VLM for remote sensing. While models such as RSGPT and RemoteCLIP focus on foundational image-text alignment, approaches like GeoChat and UniRS extend these capabilities to spatial reasoning and temporal analysis. Furthermore, innovations in model architecture, such as RSUniVLM's granularity-oriented design and SkyEyeGPT's instruction-based paradigm, underscore the growing sophistication of VLM in addressing the diverse and unique demands of remote sensing. The success of SkyEyeGPT shows the additional benefit for model training on different tasks as well as the usability improvements of a model that is able to server various tasks out of the box.

This leads us to the development of our approach by extending the training dataset as well as not only adapt the architecture but rephrasing the way data is presented to the model. Additionally, we want to explore the efficiency of smaller models to match the reached performance.

glossaries

3 Goal and Approach

The following section will describe the goal of the associated master's thesis. We will express questions that the thesis will try to answer. Additionally, we

will sketch the approach we will implement in the thesis.

3.1 Goal and Research Questions

The goal of the master's thesis is to provide an alternative approach to currently separated VL tasks and to evaluate its quality on RS data. By rephrasing the objective of different VL tasks the respective task will be transformed into a VQA task. We expect this to have a positive effect on the performance of VQA as we introduce additional question-answer pairs with supplementary context. Furthermore, by applying small language models (SLMs) we expect positive effects on general resource consumption.

Therefore, the main goals of the master's thesis can be summarized in the following research questions:

1. What is the effect of rephrasing VL tasks like IC or CC as VQA on the performance on the original task?
2. What is the effect of training VQA-models with rephrased datasets on the performance on native VQA data?

Since we are interested in reducing the footprint of fine-tuned models on VL tasks we will leverage SLMs. Therefore, we will also answer questions related to resource consumption and quality compared to large language models (LLMs). The main research question will be:

3. Can SLMs achieve competitive quality to LLMs if trained with the rephrased objective?

3.2 Approach

In order to achieve this, we will unify different datasets with specific characteristics for different VL tasks. We will then associate each dataset with task specific questions and use this strategy to rephrase each VL tasks as VQA task.

Image captioning could be associated with the following questions

- What does this image show?
- What does this image depict?
- ...

Applying this idea to a set of available tasks and datasets will lead to a generalized view on these tasks represented by question-answer pairs. After this step of rephrasing the original tasks, a Q-Former[5] based encoder-decoder model will be trained on the data.

Generally, the Q-Former[5] relies on LLMs to generate language based on the encoded prompts and images.

Instead, a small language model (SLM) can be used in place of a large language model (LLM) to handle lightweight text generation, making the model

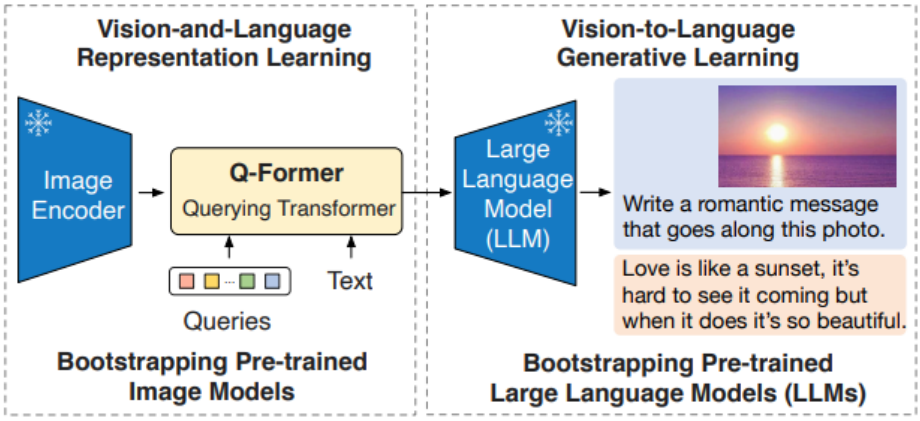

Figure 1: Q-Former architecture. Image taken from Li et al. [5].

more efficient for edge deployment and low-power devices. There are multiple benefits to it like lower resource consumption, the ability to perform inferences on edge devices or even faster inferences. As part of the master's thesis parameter-efficient models for vision and language models have to be evaluated. Language models that might be a fit are for example ALBERT[17], DistilGPT2 [6], MiniLM[18], GPT-NeoX [7]. In the thesis we will focus on Generative Pretrained Transformer (GPT)-like language models like DistilGPT2 or GPT-NeoX. Analog approaches exist for the vision encoder of the encoder-decoder architecture. Light-weight vision models are for example MobileViT[8], Efficient-Former[9] or Lite Vision Transformer [19]. We will mainly focus on Efficient-Former and MobileViT architectures.

4 Datasets and Evaluation

Once we implemented the described approach, we will train and evaluate models using different publicly available datasets that contain remote sensing images and incorporate different VL tasks. In this section we describe the used datasets and evaluation methods that will be used in the associated master's thesis.

4.1 Datasets

The sketched approach relies on publicly available RS datasets for the different VL tasks. In this section we will highlight a selection of commonly used VL tasks datasets. As part of the master's thesis we will investigate what mix of datasets will lead to a high quality of generated answers. Additionally, the mix will aim for a high degree of diversity in order to provide valuable learning signals.

For further information on the dataset candidates please refer to the AppendixA.

4.2 Evaluation

Once the described approach is implemented and models are trained using the aforementioned datasets, we need to evaluate the performance of the models. In order to judge the quality of the generated text, we will implement Bilingual Evaluation Understudy (BLEU)[20], Consensus-based Image Description Evaluation (CIDEr)[21], Semantic Propositional Image Caption Evaluation (SPICE)[22], Recall-Oriented Understudy for Gisting Evaluation (ROUGE)[23] and Metric for Evaluation of Translation with Explicit Ordering (METEOR)[24] scores as they are common metrics used to evaluate VQA approaches. For more detailed information see Appendix B. We will use these metrics to evaluate the trained models on the rephrased objective as well as metrics for the original tasks. The scores will then be used to evaluate the performance relative to expert models trained on the original task.

Besides this automated metric calculation we will also aim for a human inspection to validate the numeric results.

Procedure

To address the research questions outlined in Section 3.1 , the evaluation will proceed as follows: First, the trained model will generate predictions on a held-out subset of the dataset. These predictions will be evaluated using the automated metrics described in Section 4.2. For tasks where results can be transformed back into the original task space, additional metrics such as accuracy and F1-score will be computed. The automatically calculated metrics will serve as primary indicators of the method's effectiveness. To complement these automated assessments, manual inspection of results will be conducted on a diverse subset of the dataset to validate the derived quality scores.

To further contextualize the performance of the proposed approach, we will compare it against the reported or reproduced results of state-of-the-art methods. Additionally, we will analyze the model's performance with respect to model size to evaluate the impact of the proposed technique on computational resource requirements. This analysis aims to assess both the efficiency of fine-tuning and the environmental implications of the training method.

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. "Show and Tell: A Neural Image Caption Generator." arXiv: 1411.4555 [cs]. (Apr. 20, 2015), [Online]. Available: http://arxiv.org/abs/1411.4555 (visited on 01/26/2025), pre-published.

[2] D. H. Park, T. Darrell, and A. Rohrbach. "Robust Change Captioning." arXiv: 1901.02527 [cs]. (Apr. 17, 2019), [Online]. Available: http://arxiv.org/abs/1901.02527 (visited on 01/27/2025), pre-published.

[3] A. Agrawal, J. Lu, S. Antol, *et al.* "VQA: Visual Question Answering." arXiv: 1505.00468 [cs]. (Oct. 27, 2016), [Online]. Available: http://arxiv.org/abs/1505.00468 (visited on 01/26/2025), pre-published.

[4] R. Bommasani, D. A. Hudson, E. Adeli, *et al.* "On the Opportunities and Risks of Foundation Models." arXiv: 2108.07258 [cs]. (Jul. 12, 2022), [Online]. Available: http://arxiv.org/abs/2108.07258 (visited on 01/27/2025), pre-published.

[5] J. Li, D. Li, S. Savarese, and S. Hoi. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." arXiv: 2301.12597 [cs]. (Jun. 15, 2023), [Online]. Available: http://arxiv.org/abs/2301.12597 (visited on 01/09/2025), pre-published.

[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter." arXiv: 1910.01108 [cs]. (Mar. 1, 2020), [Online]. Available: http://arxiv.org/abs/1910.01108 (visited on 01/09/2025), pre-published.

[7] S. Black, S. Biderman, E. Hallahan, *et al.* "GPT-NeoX-20B: An Open-Source Autoregressive Language Model." arXiv: 2204.06745 [cs]. (Apr. 14, 2022), [Online]. Available: http://arxiv.org/abs/2204.06745 (visited on 01/09/2025), pre-published.

[8] S. Mehta and M. Rastegari. "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer." arXiv: 2110.02178 [cs]. (Mar. 4, 2022), [Online]. Available: http://arxiv.org/abs/2110.02178 (visited on 01/09/2025), pre-published.

[9] Y. Li, G. Yuan, Y. Wen, *et al.* "EfficientFormer: Vision Transformers at MobileNet Speed." arXiv: 2206.01191 [cs]. (Oct. 11, 2022), [Online]. Available: http://arxiv.org/abs/2206.01191 (visited on 01/09/2025), pre-published.

[10] Y. Hu, J. Yuan, C. Wen, X. Lu, and X. Li. "RSGPT: A Remote Sensing Vision Language Model and Benchmark." arXiv: 2307.15266 [cs]. (Jul. 28, 2023), [Online]. Available: http://arxiv.org/abs/2307.15266 (visited on 01/07/2025), pre-published.

[11] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan. "GeoChat: Grounded Large Vision-Language Model for Remote Sensing." arXiv: 2311.15826 [cs]. (Nov. 24, 2023), [Online]. Available: http://arxiv.org/abs/2311.15826 (visited on 01/09/2025), pre-published.

[12] F. Liu, D. Chen, Z. Guan, *et al.* "RemoteCLIP: A Vision Language Foundation Model for Remote Sensing." arXiv: 2306.11029 [cs]. (Apr. 16, 2024), [Online]. Available: http://arxiv.org/abs/2306.11029 (visited on 01/09/2025), pre-published.

[13] A. Radford, J. W. Kim, C. Hallacy, *et al.* "Learning Transferable Visual Models From Natural Language Supervision." arXiv: 2103.00020 [cs]. (Feb. 26, 2021), [Online]. Available: http://arxiv.org/abs/2103.00020 (visited on 01/09/2025), pre-published.

[14] Y. Li, W. Xu, G. Li, *et al.* "UniRS: Unifying Multi-temporal Remote Sensing Tasks through Vision Language Models." arXiv: 2412.20742 [cs]. (Dec. 30, 2024), [Online]. Available: http://arxiv.org/abs/2412.20742 (visited on 01/09/2025), pre-published.

[15] X. Liu and Z. Lian. "RSUniVLM: A Unified Vision Language Model for Remote Sensing via Granularity-oriented Mixture of Experts." arXiv: 2412.05679 [cs]. (Dec. 10, 2024), [Online]. Available: http://arxiv.org/abs/2412.05679 (visited on 01/09/2025), pre-published.

[16] Y. Zhan, Z. Xiong, and Y. Yuan. "SkyEyeGPT: Unifying Remote Sensing Vision-Language Tasks via Instruction Tuning with Large Language Model." arXiv: 2401.09712 [cs]. (Jan. 18, 2024), [Online]. Available: http://arxiv.org/abs/2401.09712 (visited on 01/09/2025), pre-published.

[17] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." arXiv: 1909.11942 [cs]. (Feb. 9, 2020), [Online]. Available: http://arxiv.org/abs/1909.11942 (visited on 01/09/2025), pre-published.

[18] 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." arXiv: 2002.10957 [cs]. (Apr. 6, 2020), [Online]. Available: http://arxiv.org/abs/2002.10957 (visited on 01/09/2025), pre-published.

[19] C. Yang, Y. Wang, J. Zhang, *et al.* "Lite Vision Transformer with Enhanced Self-Attention." arXiv: 2112.10809 [cs]. (Dec. 20, 2021), [Online]. Available: http://arxiv.org/abs/2112.10809 (visited on 01/09/2025), pre-published.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds., Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. [Online]. Available: https://aclanthology.org/P02-1040/ (visited on 01/07/2025).

[21] R. Vedantam, C. L. Zitnick, and D. Parikh. "CIDEr: Consensus-based Image Description Evaluation." arXiv: 1411.5726 [cs]. (Jun. 3, 2015), [Online]. Available: http://arxiv.org/abs/1411.5726 (visited on 01/07/2025), pre-published.

[22] P. Anderson, B. Fernando, M. Johnson, and S. Gould. "SPICE: Semantic Propositional Image Caption Evaluation." arXiv: 1607.08822 [cs]. (Jul. 29, 2016), [Online]. Available: http://arxiv.org/abs/1607.08822 (visited on 01/07/2025), pre-published.

[23] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013/ (visited on 01/07/2025).

[24] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds., Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909/ (visited on 01/07/2025).

[25] Q. Cheng, H. Huang, Y. Xu, Y. Zhou, H. Li, and Z. Wang, "NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022, ISSN: 1558-0644. DOI: 10.1109/TGRS.2022.3201474. [Online]. Available: https://ieeexplore.ieee.org/document/9866055 (visited on 01/07/2025).

[26] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*, Jul. 2016, pp. 1–5. DOI: 10.1109/CITS.2016.7546397. [Online]. Available: https://ieeexplore.ieee.org/document/7546397 (visited on 01/07/2025).

[27] Z. Yuan, W. Zhang, K. Fu, *et al.*, "Exploring a Fine-Grained Multiscale Method for Cross-Modal Remote Sensing Image Retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022, ISSN: 1558-0644. DOI: 10.1109/TGRS.2021.3078451. [Online]. Available: https://ieeexplore.ieee.org/document/9437331 (visited on 01/07/2025).

[28] 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 Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–23, 2024, ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2024.3449154. arXiv: 2306.11300 [cs]. [Online]. Available: http://arxiv.org/abs/2306.11300 (visited on 01/07/2025).

[29] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring Models and Data for Remote Sensing Image Caption Generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018, ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2017.2776321. arXiv: 1712.07835 [cs]. [Online]. Available: http://arxiv.org/abs/1712.07835 (visited on 01/07/2025).

[30] G. Cheng, J. Han, and X. Lu, "Remote Sensing Image Scene Classification: Benchmark and State of the Art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017, ISSN: 0018-9219, 1558-2256. DOI: 10.1109/JPROC.2017.2675998. arXiv: 1703.00121 [cs]. [Online]. Available: http://arxiv.org/abs/1703.00121 (visited on 01/07/2025).

[31] C. Liu, R. Zhao, H. Chen, Z. Zou, and Z. Shi, "Remote Sensing Image Change Captioning With Dual-Branch Transformers: A New Method and a Large Scale Dataset," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022, ISSN: 1558-0644. DOI: 10.1109/TGRS.2022.3218921. [Online]. Available: https://ieeexplore.ieee.org/document/9934924 (visited on 01/07/2025).

[32] Y. Zhan, Z. Xiong, and Y. Yuan, "RSVG: Exploring Data and Models for Visual Grounding on Remote Sensing Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023, ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2023.3250471. arXiv: 2210.12634 [cs]. [Online]. Available: http://arxiv.org/abs/2210.12634 (visited on 01/07/2025).

[33] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual Question Answering for Remote Sensing Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8555–8566, Dec. 2020, ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2020.2988782. arXiv: 2003.07333 [cs]. [Online]. Available: http://arxiv.org/abs/2003.07333 (visited on 01/07/2025).

[34] S. Lobry, B. Demir, and D. Tuia, "RSVQA Meets Bigearthnet: A New, Large-Scale, Visual Question Answering Dataset for Remote Sensing," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, Jul. 2021, pp. 1218–1221. DOI: 10.1109/IGARSS47720.2021.9553307. [Online]. Available: https://ieeexplore.ieee.org/document/9553307 (visited on 01/07/2025).

[35] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. Murphy. "FloodNet: A High Resolution Aerial Imagery Dataset for Post Flood Scene Understanding." arXiv: 2012.02951 [cs]. (Dec. 5, 2020), [Online]. Available: http://arxiv.org/abs/2012.02951 (visited on 01/07/2025), pre-published.

A Datasets

Image Captioning

There are various datasets that target IC in RS e.g. [25]–[29]. However, as RS5M [28] and RSCID [29] will provide us with a large and divers collection of IC data, we will mainly focus on these datasets to represent the IC task.

RS5M[28] is a large-scale vision-language dataset specifically designed for remote sensing applications, containing 5 million image-text pairs. The dataset was constructed by combining publicly available image-text datasets with labeled remote sensing datasets, where captions were generated using pre-trained VLMs. The images cover a wide range of remote sensing contexts, including urban, agricultural, forest and aquatic environments, while the corresponding textual descriptions provide rich and detailed annotations of the scene content, emphasizing geographical and environmental characteristics. RS5M was curated to ensure high-quality data, removing noise and irrelevant entries.

RSCID[29] is designed to facilitate research in generating descriptive captions for remote sensing images. The dataset comprises 10,921 images, each accompanied by five human-annotated captions, resulting in a total of 54,605 descriptions. The images cover a wide range of scenes, including urban, rural and natural environments, with diverse objects and structures.

Scene Classification

NWPU-RESISC45[30] dataset is a large-scale benchmark designed for remote sensing image scene classification, introduced by Cheng et al. in 2017. This dataset comprises 31,500 images distributed across 45 scene classes with 700 images per class. Each image has a resolution of 256x256 pixels. The images were collected from over 100 countries. NWPU-RESISC45 contains variations in translation, spatial resolution (ranging from 0.2 to 30 meters per pixel), viewpoint, object pose, illumination, background and occlusion.

Change Captioning

LEVIR-CC[31] is a large-scale benchmark designed for RS CC. The dataset comprises 10,077 image pairs, each capturing the same geographic region at two different times and includes 50,385 human-written captions describing the observed changes. These captions highlight changes, such as new construction, vegetation growth, or land-use modifications. Each image in the dataset has a resolution of 256x256 pixels. The captions are curated to ensure high quality and diversity, with multiple descriptions per image pair to account for linguistic variability. LEVIR-CC serves as a valuable resource for training and evaluating

models for vision-language tasks, particularly in remote sensing and supports advancements in change detection and multimodal learning. The dataset is publicly available, fostering further research in this emerging field.

Visual Grounding

DIOR_RSVG[32] is a large-scale benchmark designed to advance research in Remote Sensing Visual Grounding (RSVG). This dataset comprises 38,320 image-expression-box triplets, derived from 17,402 remote sensing images, where each triplet includes an image, a natural language expression and a corresponding bounding box that localizes the object referred to in the expression. The average length of the expressions is 7.47 words, providing detailed descriptions for precise object localization.

Visual Question Answering

RSVQA[33] This paper introduces the RSVG Dataset (RSVGD), a large-scale benchmark for Remote Sensing Visual Grounding (RSVG). The dataset comprises image-expression-box triplets, where each triplet includes a remote sensing image, a natural language expression and a bounding box indicating the object referred to in the expression. RSVGD encompasses a diverse range of object categories and scenes, reflecting the complexity of remote sensing imagery. The dataset is publicly available, facilitating research in visual grounding within the remote sensing domain.

RSVQA x BEN[34] This work presents the RSVQAxBEN dataset, a large-scale Visual Question Answering (VQA) dataset tailored for remote sensing applications. Derived from the BigEarthNet dataset, RSVQAxBEN contains approximately 15 million image-question-answer triplets. Each triplet consists of a Sentinel-2 image, a question formulated in natural language and the corresponding answer. The dataset supports various question types, including scene classification, object counting and attribute recognition, making it a valuable resource for developing and evaluating VQA models in the context of remote sensing.

FloodNet[35] The FloodNet dataset comprises high-resolution Unmanned Aerial Vehicle (UAV) imagery captured after Hurricane Harvey, focusing on post-flood scene understanding. It includes 2,343 images with pixel-wise annotations for semantic segmentation tasks, covering categories such as flooded buildings, non-flooded buildings, roads and water. Additionally, FloodNet offers approximately 11,000 question-image pairs for Visual Question Answering (VQA), addressing challenges like detecting flooded infrastructure and distinguishing between natural water bodies and floodwater.

Mixed Task Datasets

RSICap and RSIEval have been introduced by Hu et al.[10]. The RSICap is a curated collection comprising 2,585 high-quality, human-annotated captions tailored to the remote sensing domain. Each image in the dataset is paired with detailed descriptions that encapsulate both scene-level information, such as residential areas, airports and farmlands as well as object-specific attributes including color, shape, quantity and absolute position. The RSIEval includes human-annotated captions alongside visual question-answer pairs. The QA pairs span various question types, including object-centric queries, spatial relationship questions and scene-level queries . These pairs challenge models to interpret intricate visual details, reason about spatial arrangements and contextualize the broader scene.

B Metrics

To assess the quality of the generated text we will use BLEU, ROUGE, METEOR, CIDEr and SPICE as metrics depending on the task.

Automated Evaluation

In addition to these metrics dedicated to language generation we will also report Accuracy, Recall and F1-Score for original classification tasks.

1. **BLEU**[20] measures n-gram overlap to assess syntactic similarity between generated and reference answers. While it is sensitive to phrasing variations, it provides a useful baseline for lexical alignment.

2. **ROUGE** evaluates n-gram recall, ensuring key phrases and concepts from the reference are adequately captured in the generated response.

3. **METEOR**[24] improves upon BLEU by accounting for synonyms, stemming, and word order. This makes it effective for identifying semantically equivalent but differently phrased answers.

4. **CIDEr**[21] utilizes TF-IDF weighting of n-grams to evaluate semantic relevance across multiple reference answers. It is particularly suitable for tasks involving diverse linguistic expressions.

5. **SPICE**[22] focuses on assessing semantic and relational content using a scene-graph-based evaluation, making it particularly valuable for long descriptive answers.

Human Evaluation

While automated metrics provide efficiency and consistency, they may fail to capture subtleties such as fluency, coherence and domain relevance in responses.

To address this, human evaluation will be conducted on a subset of the data, focusing on:

- **Semantic Alignment** The degree to which the generated answer correctly addresses the question based on the visual context.
- **Descriptive Quality** The richness, clarity and informativeness of the response.
- **Coherence** The logical flow and consistency of the answer, particularly for multi-sentence responses.