Remote Sensing Image Caption Generation Using the RSICD Dataset

Jason Widjaja z549497
Revathi Sridhar Surya z5520480
Shayan Ziaei z1234567
Sachin Singh z5525041
Pankaj Patel z5519861

Abstract
Remote Sensing Image Captioning (RSIC) is the task of generating captions from remote sensing images, which could be applied to urban planning, agricultural monitoring and disaster response, etc. RSIC, in contrast, is particularly challenging because the object scales are arbitrary and the textures appear visually similar, even though they convey different meanings. In this paper, we explore both of the improvements on the RSICD dataset: first, geometric and photometric augmentation with rare-word prioritisation, and second, model capacity scaling from BLIP1 Base to BLIP1 Large. Although data augmentation improved result metrics, it permitted models to scale to BLIP1 Large, making significantly more coherent and descriptive captions with better spatial reasoning. As a result, our findings show that increasing model capacity and thoughtful evaluation of data strategies are key to developing and advancing RSIC performance.

Keywords: Blip, RSICD, Augmentation, Metrics, Caption

1. INTRODUCTION
Image captioning is the task of generating natural language descriptions from visual data. While extensively studied for natural images (e.g., COCO, Flickr), applying these techniques to remote sensing imagery—such as satellite or aerial views—remains significantly more challenging. These images often contain objects at vastly different scales, lack a fixed orientation, and exhibit visually similar textures that correspond to semantically distinct classes (e.g., meadows vs. farmland). A single image may include roads, rivers, buildings, and vegetation, making scene understanding and caption generation inherently complex. Automatically generating captions for such imagery enables high-level, interpretable summaries of land use, infrastructure, and environmental conditions. This has real-world value in domains such as urban planning, agriculture, disaster response, and environmental monitoring—especially for rapid assessment or large-scale geographic analysis.

2. RELATED WORK
The newly emerging task of Remote sensing image captioning (RSIC) tries to fill this semantic gap between complex aerial imagery and natural language descriptions. The following articles represent breakthroughs in the evolution of RSIC models, advancing from traditional encoder-decoder structures to more modernised architectures involving attention-based modules and transformers.

2.1 LITERATURE REVIEW
Foundations with Generic Attention (2017)
Lu et al. (2017) — RSICD: A Benchmark Database for Remote Sensing Image Captioning with Three Pilot Tasks. This model was the first who introduced the RSIC problem, where they introduced an RSIC dataset (with over 10k images and five descriptions for each image) and they constructed CNN-LSTM frameworks using Adaptive attention based on Show Attend Tell in their research to run ROI(CNN: AlexNet, VGG GoogLeNet). a multimodal encoder-decoder net without attention achieved BLEU-4: 0.27 and CIDEr: 2.03; however, in the presence of either soft or hard attention it will outperform this (BLEU-4: 0.37 and CIDEr: 2.02 on RSICD)., however it was limited by simple sentences and poor generalisation across different datasets, low-quality datasets due to sentence repetition, and remote sensing specific issues scale/rotation ambiguity complex spatial relationships.

Structured and Label-Guided Attention (2019–2021)
[2] Zhang et al. A Label-Attention Mechanism (LAM) had been proposed by Chen et al., 2019, where the attention was based on label words of remote sensing image classification to the semantic segments. VGG16 + LSTM decoder 4.7 BLEU-4 and 19.7 CIDEr reported on RSICD (vs BLEU-4:0.403, CIDEr:2.585 by their model). Meanwhile, [3] Zhao et al. Structured Attention (2021), with this time, semantic segmentation as the base training task for structured attention maps that are conditioned on image regions.BLEU-4:0.448 CIDEr:1.037 This step marks the first step towards attention about not pure pixels but semantic and spatial interpretable patterns. Single-label classification cannot explain more than one kind of scene mixed into an image, and model prediction also depends on the labelling quality

Multi-Level and Multi-Scale Attention (2020–2023)
[4] Li et al. Vinokurov et al. (2020) proposed a multi-level attention mechanism to consider attention over image regions and decoded words at the same time, as well as vision-semantic balance for human-like attention combination. Although their model achieved BLEU-4: 0.516 and CIDEr: 2.77 in the RSICD data set, it had quality of data sets (error rate is 13,12%) and computational cost problems. [5] Li et al. (2024) expanded on this with VSCA-Net, performing CIDEr 2.93 of visual-semantic embeddings [6], Cheng et al. VIFAP (w/o 2022) fused with MLS-Net (CIDEr: 2.36), [7] Zhang et al. (2023): MSISAM (CIDEr: 2.84) [37] proposed a hierarchical stair attention. Models began to generalise more slowly to objects at different sizes and scenes with greater complexity; however, the quality of the datasets remained a bottleneck in addition to scalability issues when designing multi-level architectures.

Transformer Architectures and Semantic Enhancement (2022–2024)
[8] Wang et al. Although CapFormer is the first solely pure transformer model of RSIC, it retained and replaced LSTMs with a ViT encoder and transformer decoder, unlike(2022). It achieved a CIDEr: 3.15, suggesting that global self-attention had the potential of capturing semantically dependent information across spatial layouts. At the same time. In parallel, [9] Liu et al. Network (2022), which uses MLAT that extracts features from

multiple CNN layers, and decodes with a transformer to achieve ~2.8 CIDEr. Later, [10] Meng et al. (2023, 2024): PKG-TR: Transformer, MG-TR: Transformer (scene-object semantic embeddings using CLIP, up to 3.2 CIDEr). These models demonstrated that transformer-based world context modelling methods and multi-modal semantic alignment can be applied to generate the response, and achieved a new performance in fluency, coherence, and descriptiveness of RSIC. Despite all this architectural progress, however, most still hinge on pre-trained CNN features and proposals or CLIP embeddings — and are therefore vulnerable to upstream errors in these steps, including any biases from natural images.

Meta-Learning, Dual Tasks, and Region-Aware Transformers (2024–2025)

Yang et al. (2022) used meta-learning by training on auxiliary classification tasks to build a universal encoder, BLEU-4: 0.680, CIDEr: 2.742 (Meta-Caption). [12] Ye et al. JTTS (2022): Two-stage method first predicting attribute vectors before captioning, enhanced with semantic knowledge via an attribute-guided transformer decoder (CIDEr: ~2.8). [13] Guo et al. Next, CCT (Cooperative Connection Transformer) (2024), which integrated features of grid-level and region-level for better focusing and less noise, with an average CIDEr score of 3.07. Finally, [14] Zhao et al. (2024) Region Attention Transformer (RAT), shepherds region-level attention using RoIs and benchmarks with region annotations (CIDEr: ~2.9). These use cases are based on sample efficiency driven by spatial structure but still limitedly and dependent on the downstream classifier, region proposal alternation, or auxiliary attribute tagging.

3. METHOD

This project aims to describe remote sensing images using natural language captions by implementing deep learning architectures. We tested three baseline models: (LAM) Label-Attention-Mechanism, a mix of Vision Transformer and GPT2 and BLIP-1(Bootstrapping-Language-Image Pretraining). The models were trained and tested on the RSICD dataset, a collection of 10,921 high-resolution satellite images with human-crowd-sourced image captions. To increase the transfer capability of each architecture, we also fine-tuned pretrained modules. A straightforward comparison of performance was made based on traditional captioning metrics: BLEU (1–4), METEOR, ROUGE-L, CIDEr, and SPICE.

3.1 Label-Attention Mechanism (LAM)

We use the [15]LAM Label-Attention-Mechanism, which is composed of an encoder-decoder architecture for image captioning using spatial features extracted by pre-trained VGG16, a small linear function providing a classification of the image and an LSTM as a decoder to generate the caption with label-based attention. The Attention Mechanism in LAM is Label-guided Attention, which better utilises this feature over the image. The model part consists of the Image Feature Path, the Hidden State Path, and the Label-Attention Path, which are merged for relevance computation. Unlike concatenation used in baseline pipelines. This would be beneficial as it would be less category-agnostic and focus well on the more information-rich regions. Due to this, the decoder is more adept at articulating complex manifolds as well as high-frequency details. In addition, the direct association between semantic label cues and visual attention in our LAM leads to more coherent captions.

3.2 Proposed Model: Vision Transformer (ViT) +GPT-2

The proposed Model combines a Vision Transformer (ViT) for image encoder and GPT-2 as decoder for the caption generation of the RSIC Dataset. The ViT processes the input image by splitting it into smaller sections and applying self-attention mechanisms to capture how different areas relate to each other, producing an overall visual representation. A transformer-based language model pre-trained on massive text corpora, word by word. ViT has a powerful visual comprehension and, through GPT-2 a fluent language generation, to achieve context-sensitive and semantically dense captions. The model is learned end-to-end, and no additional labels are needed for segmentation maps or handcrafted features, making it scalable and effective for diverse remote sensing imagery.

3.3 BLIP(BootstrappingLanguage Image Pretraining)

BLIP is intended to be a universal vision-language model suitable for both comprehension and generation tasks. Instead, it uses a Vision Transformer (ViT) to convert images into patches while learning one global representation. For text, it uses a BERT-like encoder and a causal language model decoder. A key ingredient of BLIP is the Multimodal Mixture of Encoder-Decoder (MED) architecture, which uses shared vision-and-language transformer layers and cross-attention. The architecture enables BLIP to reliably condition on both visual and textual information, guiding the model to learn how images should be interpreted and what captions should make sense in response. The design of BLIP is efficient but open-ended, providing a strong baseline for transferring the recent advances in general vision-language models to the remote sensing domain.

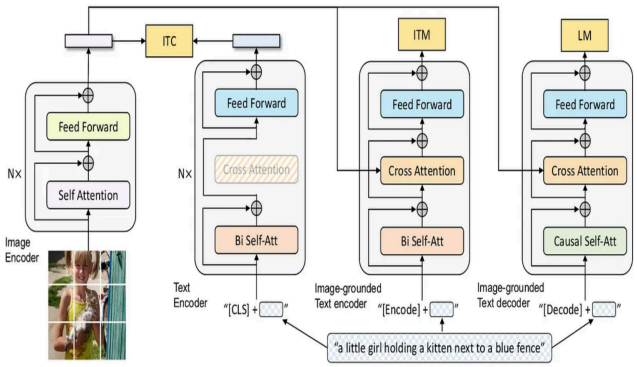

Figure 1. BLIP (Bootstrapping Language-Image Pretraining)

Models	Blue 1	Blue 2	Blue 3	Blue 4	Meteor	Rouge	Cider	Spice
LAM	4.6	1.78	0.98	0.62	9.8	9.67	-	-
VIT + GPT	58.32	34.56	21.18	13.71	34.13	33.06	38.46	21.24
BLIP 1 Base	68.09	50.71	38.65	30.27	25.79	47.94	58.64	26.71

Table 1. Base model ResultS

As we can see in Table 1, BLIP 1 outperforms LAM and Vit + GPT and was therefore selected as our final proposed approach.

3.4 Fine Tuning

Data Augmentation :

For data augmentation, we used different types of transformation techniques on the RSICD dataset to expand the data as (90/270) rotation, horizontal flip, jittered intensity change, contrast change, sharpen and light change..Similarly, the ITF was inverted and rare words in captions were automatically given more weight, such that training set frequencies of these words were tripled. The techniques expanded the dataset from 43,680 to 131,040 caption–image pairs. The augmented techniques model has shown significant improvements in all metrics, with a 14.5% gain in CIDEr, and increases in both Meteor and ROUGE-L. These results were strongest for BLEU-1 and BLEU-2, which were better than other metrics, whereas BLEU-4 and SPICE performed worse than the mean. Augmentation positively affected coverage and relevance, but there was not much difference in terms of syntactic diversity.

Upgrade Architecture:

The captioning model improved its visual understanding and language generation by upgrading from BLIP1 Base to BLIP1 Large. Extending Vision Encoder: It upgrades the image encoders from ViT-B/16(86M) to a larger version like ViT-L/16(307M) for the above capability. Image Encoder: It was improved from the common BERT-base (124M) to an even bigger BERT-large (340M), allowing it a much larger vocabulary, since that is what makes the text decoding more difficult. So now you could have a lot of different word variations once combined to create sentences (if they are semantically and grammatically correct)..The 210M parameters were increased to 647M, and the embedding dimension was increased from 768 to 1024, resulting in a ~33% increase in representational capacity for better spatial reasoning (e.g. "airports near highways"). This was done by increasing the number of attention heads from 12 to 16 for improved parallel spatial reasoning and disambiguation of objects, as well as layers from 12 to 24 for deeper semantic reasoning in addition to more descriptive captions. However, while the model had been improved in this manner, it still lacked terms for remote sensing that were sufficiently tailored to take advantage of these domain-specific semantic constraints.

4. EXPERIMENTS

4.1 Data Analysis

4.1.1 Dataset Overview

The Remote Sensing Image Captioning Dataset (RSICD) is a large-scale benchmark established in an end-to-end manner to support research on automatic descriptions for satellite and aerial imagery. It consists of 10,921 remote sensing images, and each image has five human-written descriptions that form a total of 54,605 image-sentence pairs. It is collected from different geographic sources (Google Earth, Baidu Map, MapABC, Tianditu), which provide diverse land cover conditions, climates, acquisition angles and spatial resolutions.

Example:

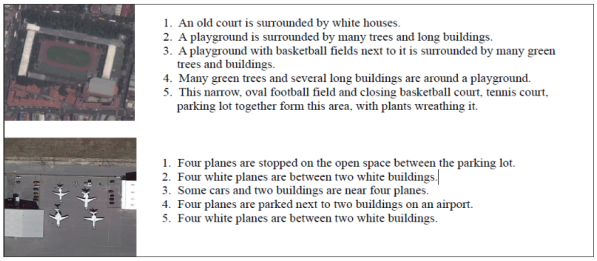

Figure 2.: The example of images and corresponding five sentences, each image selected from our dataset.

- **Dataset URL**: https://github.com/201528014227051/RSICD_optimal

4.1.2 Dataset Properties and Preprocessing

Each image in RSICD has a size of 224×224 dimensions, in RGB colour mode, which is compatible with convolutional and transformer-based deep learning models. In contrast, the Blip 1 model resizes the image to 384 x 384 dimensions to match the model input resolution.

*Resolution: All images are of size 224×224 in the dataset
* Format : RGB images
* Total Caption: 5 per image, Describing visible content

4.1.3 Class Categories and Distribution

The RSICD dataset has 31 categories, which are divided into 3 groups: urban infrastructure (airports, stadiums, bridges, residential and commercial areas), natural landscape (forests, rivers, farmland, mountains, beaches) and leisure buildings (play fields, churches, palaces, squares). Each image comes with 5 human-annotated descriptions and a detailed explanation of the content, such as infrastructure (e.g. man-made features), vegetation (e.g. plants, trees) and land usage (e.g. agriculture, residential) information visible in the image.

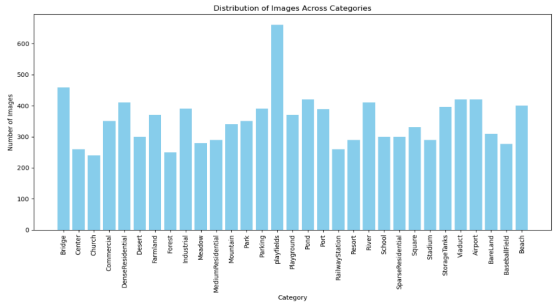

Figure 3. Data Distribution

4.1.4 Challenges in the Dataset

Class Imbalance: The imbalance of the class representations (as seen in Figure 2) can bias the model to produce captions that overfit to the common scene types. This has to be handled by means such as weighted sampling, loss rebalancing, and data augmentation.

Complexity of the scene: The images tend to have more than one non-overlapping object or land use types (e.g. road passing through residential and commercial zones) – caption generation can be quite complicated.

Semantic Vagueness: Visually similar Textures (e.g., farmland vs. meadow, forest vs. orchard) make the model hard to generate semantically accurate descriptions.

Orientation ambiguity: The images in the remote sensing domain may have arbitrary orientations (north need not be on top), which makes the tasks of object recognition and spatial understanding difficult for the network.

4.2 Evaluation Strategy

Along with the RSIC data set, we evaluated five traditional natural language generation metrics: BLEU(1–4), METEOR, ROUGE-L, CIDEr, and SPICE for our image captioning models. BLEU measures n-gram accuracy of a linguist between the reference caption and the generated caption, and BLEU-1 through 4 score higher lengths of phrases. While METEOR is a BLEU-like with major differences, it enhances the synonym match

solutions and word order. ROUGE-L scores the longest common subsequence between reference and generated captions, thus making it more suitable for estimating sequence similarity (fluency). Specifically, as CIDEr was proposed for image captioning to compare human and machine captions of an image, SPICE evaluates on semantic scene understanding instead of surface word co-occurrence. These all metrics help to evaluate the Caption generated from the image for all three models.

4.3 Hyperparameters:

For all our experiments, we use an identical training set when comparing to BLIP 1. We mainly focus on analysing the performance of BLIP 1, chosen as a proposed model, and find out that using optimised hyperparameters to further finetune it can significantly sharpen the generated captions on the RSIC dataset. We used Adam optimiser with a learning rate of 2e-5 for training, the batch size was fixed to 16, and we avoided unnecessary epochs using early stopping.

5. Results

Our experiments on the RSIC dataset were evaluated with two improvement strategies: data augmentation, architectural scaling, as shown in the following tables.

Models	Blue 1	Blue 2	Blue 3	Blue 4	Meteor	Rouge	Cider	Spice
BLIP 1 Base	68.09	50.71	38.65	30.27	25.79	47.94	58.64	26.71
AFTER	64.45	51.31	38.57	29.60	26.63	48.81	67.17	25.20

Table 2. Result After Augmentation

Table 1 compares the performance of the BLIP-1 Base model with the AFTER model across multiple metrics. The AFTER model shows a slight decrease in Blue1 (64.45 vs. 68.09) but achieves improvements in Cider (67.17 vs. 58.64) and comparable performance across Blue2, Blue3, Blue4, Meteor, and Rouge. However, Spice drops slightly from 26.71 to 25.20.

Models	Blue 1	Blue 2	Blue 3	Blue 4	Meteor	Rouge	Cider	Spice
BLIP 1 Base	68.09	50.71	38.65	30.27	25.79	47.94	58.64	26.71
BLIP1 Large	73.87	57.73	45.84	37.19	29.99	53.97	88.22	29.17

Table 3. Result after Upgrade Architecture

Table 2 presents the comparison between BLIP-1 Base and BLIP-1 Large. The BLIP-1 Large model outperforms the base model across all evaluation metrics, showing notable gains in Blue-1 (73.87 vs. 68.09), Blue-2 (57.73 vs. 50.71), Blue-3 (45.84 vs. 38.65), and Blue-4 (37.19 vs. 30.27). The Cider score shows the most significant improvement, increasing from 58.64 to 88.22, indicating a substantial enhancement in caption quality.

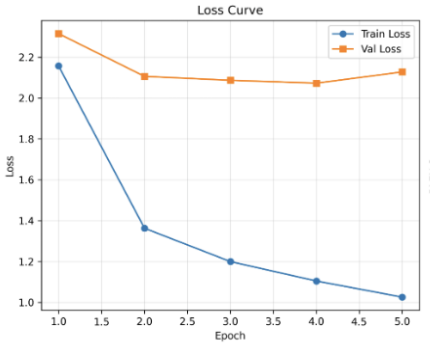

Figure 3 Training Loss vs Validation Loss

Figure 3 illustrates the training and validation loss curves over five epochs. The training loss shows a consistent and substantial decrease from approximately 2.2 in the first epoch to nearly 1.0 by the fifth epoch, indicating effective model learning. The validation loss decreases initially from about 2.25 to just above 2.0, but then exhibits a slight upward trend after the third epoch, suggesting possible overfitting in later epochs.

Figure 4 shows an example of a generated caption compared to the ground truth descriptions.

Ground Truth

1. at almost every gate of terminal building, there is a airplane .
2.some cars are near a building on an airport .
3. many planes are parked next to a building on an airport .
4. many planes are around a long building .
5. many planes are parked next to a building on an airport .

Figure 4

Generated Caption:

Some planes are parked near several buildings in an airport.

The model correctly identifies the presence of multiple planes and their location near buildings within an airport, producing the caption "Some planes are parked near several buildings in an airport." This matches the key objects and relationships described in the ground truth, such as "planes," "buildings," and "airport," demonstrating the model's ability to capture the main scene accurately, even if the wording differs from the references.

6. Conclusion

We demonstrated that our data augmentation method can improve the performance of rating-based and content-based metrics with minimal scaling up of any of the original problems in the RSICD dataset, such as repetition, limited presence, or class imbalance. Though the augmentation helped to produce more fluent and high-quality captions, it also failed to capture some domain-specific vocabulary and concepts. An updated version model BLIP large, which does better overall and produces generally more fluent, more meaningful captions with improved adjective-noun and spatial reasoning. In summary, the results indicate that there still may be a future where we should focus on making training data bigger by adding more words that are specific to RSIC and improving the model's ability to understand objects arranged in an image. This improvement will help the system to describe images more accurately and be useful for city planning, self-driving systems and disaster help

REFERENCES

[1] Lu, X., Wang, B., Zhang, J., & Zheng, X. (2017). *Exploring Models and Data for Remote Sensing Image Caption Generation*. arXiv preprint arXiv:1712.07835. https://arxiv.org/abs/1712.07835

[2] Zhang, Z.; Diao, W.; Zhang, W.; Yan, M.; Gao, X.; Sun, X. (2019).
LAM: Remote Sensing Image Captioning with Label-Attention Mechanism. Remote Sensing, 11(20), 2349.
https://doi.org/10.3390/rs11202349

[3] Zhao, R.; Shi, Z.; Zou, Z. (2021).
High-resolution remote sensing image captioning based on structured attention.
IEEE Transactions on Geoscience and Remote Sensing, 60, 5603814.
https://doi.org/10.1109/TGRS.2021.3095166

[4] Li, Y.; Fang, S.; Jiao, L.; Liu, R.; Shang, R. (2020).
A Multi-Level Attention Model for Remote Sensing Image Captions. Remote Sens. 2020, 12(6), 939.
https://doi.org/10.3390/rs12060939

[5] Li, Y.; Zhang, X.; Cheng, X.; Tang, X.; Jiao, L. (2024).
Learning Consensus-Aware Semantic Knowledge for Remote Sensing Image Captioning. Pattern Recognit. 2024, 145, 109893. https://doi.org/10.1016/j.patcog.2023.109893

[6] Cheng, Q.; Huang, H.; Xu, Y.; Zhou, Y.; Li, H.; Wang, Z. (2022).
NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning. IEEE Trans. Geosci. Remote Sens. 2022, 60, 5629419. https://doi.org/10.1109/TGRS.2022.3201474

[7] Zhang, X.; Li, Y.; Wang, X.; Liu, F.; Wu, Z.; Cheng, X.; Jiao, L. (2023).
Multi-Source Interactive Stair Attention for Remote Sensing Image Captioning. Remote Sens. 2023, 15, 579.
https://doi.org/10.3390/rs15030579

[8] Wang, J.; Chen, Z.; Ma, A.; Zhong, Y. (2022). *CapFormer: Pure Transformer for Remote Sensing Image Caption*. In Proceedings of the IEEE IGARSS 2022, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 7996–7999. [IEEE IGARSS Proceedings]

[9] Liu, C.; Zhao, R.; Shi, Z. (2022).
Remote-Sensing Image Captioning Based on Multilayer Aggregated Transformer. IEEE Geosci. Remote Sens. Lett., 19, 6506605. https://doi.org/10.1109/LGRS.2022.3149589

[10] Meng, L.; Wang, J.; Yang, Y.; Xiao, L. (2023).
Prior Knowledge-Guided Transformer for Remote Sensing Image Captioning. IEEE Trans. Geosci. Remote Sens., 61, 4706213. https://doi.org/10.1109/TGRS.2023.3240326

[11] Meng, L.; Wang, J.; Meng, R.; Yang, Y.; Xiao, L. (2024).
Multiscale Grouping Transformer With CLIP Latents for RSIC.
IEEE Trans. Geosci. Remote Sens., 62, 4703515.
https://doi.org/10.1109/TGRS.2024.3367939

[12] Yang, Q.; Ni, Z.; Ren, P. (2022).
Meta Captioning: A Meta Learning Based Remote Sensing Image Captioning Framework. ISPRS J. Photogramm. Remote Sens., 186, 190–200.
https://doi.org/10.1016/j.isprsjprs.2022.02.012

[13] Ye, X. et al. (2022).
A Joint-Training Two-Stage Method for Remote Sensing Image Captioning. IEEE Trans. Geosci. Remote Sens., 60, 4709616.
https://doi.org/10.1109/TGRS.2022.3224244

[14] Guo, W. et al. (2024).
Cooperative Connection Transformer for RSIC.

[15] Zhang, W. et al. (2019)
LAM: Remote Sensing Image Captioning with Label-Attention Mechanism. *Remote Sens.*, 11(20), 2349
https://doi.org/10.3390/rs11202349

[16] Zhao, Y. et al. (2024).
Region Attention Transformer with Region-Annotated UCM/Sydney Datasets.
Lin, H., Hong, D., Ge, S., Luo, C., Jiang, K., Jin, H., & Wen, C. (2024). *RS-MoE: A Vision-Language Model with Mixture of Experts for Remote Sensing Image Captioning and Visual Question Answering*. ArXiv.org.
https://arxiv.org/abs/2411.01595