Ensemble Captioning of Images via Voting-Based Fusion Strategies

K.Rajani[1], J.Surya Teja[2], Sk.Shajahan[3], C.Sudheer[4] V.Srinivas[5], K.Syamala Devi[6], S.Rizwana[7]
[1]Associate Professor, [2,3,4] B.Tech Students
[1,2,3,4,5,7]Department of Computer Science and Engineering,
[6]Department of Humanities and Mathematics,
[1,2,3,4,7]Narasaraopeta Engineering College, Narasaraopet, Andhra Pradesh, India
[5]GRIET, Hyderabad, Telangana, India.
[6]G. Narayanamma Institute of Technology and Science(women), Shaikpet, Hyderabad, Telangana, India
[1]rajani.kadiyala@gmail.com, [2]jadamsuryateja@gmail.com,
[3]shaikshajahan8705@gmail.com, [4]chaitanyasudheer033@gmail.com
[5]srinivassai1549@gmail.com , [6]syamaladevi@gnits.ac.in
[7]syedrizwananrt@gmail.com

***Abstract*—When seen vertically, satellite-based imagery presents difficulties for automatic caption creation. Semantic interpretation is made more difficult by the absence of intuitive indicators like perspective depth and object occlusion that are present in natural settings. Here, we develop a modular captioning pipeline that combines the outputs of two contemporary vision-language models, Kosmos-2 and BLIP-base. Instead of depending on a single caption generator, each model generates textual descriptions on its own. These descriptions are then examined using a simple vote method to determine which phrase is the most representative. The RSICD benchmark is used to test the pipeline, and standard metrics like BLEU, METEOR, ROUGE-L, CIDEr, and SPICE are used to assess it. Compared to individual models, our approach consistently produces captions that are more accurate, logical, and contextually appropriate for remote sensing scenarios.**

***Index Terms*—Remote sensing, image captioning, ensemble models, vision-language processing, fusion strategy, BLIP, Kosmos-2, RSICD.**

I. Introduction

A unique and top-down view of the Earth's surface is provided by satellite and aerial platform images, which are now crucial for landscape analysis, environmental change monitoring, and decision support in fields like agriculture, disaster relief, and urban development [1]–[4]. Even with their significance, automatically deciphering these pictures is still a challenging endeavor. Remote sensing photos lack distinct visual clues such perspective lines, shading, and depth, in contrast to traditional photographs. Rather, they show scenes that are flattened and contain pieces that are consistently organized or closely packed, which makes it more difficult for models to derive meaningful semantics.

Image captioning has traditionally advanced through encoder-decoder pipelines, which combine language models such as recurrent neural networks (RNNs) with convolutional neural networks (CNNs) to produce descriptive text [5], [6]. These solutions were mostly created for commonplace photos and general-purpose datasets like Flickr30k and MS-COCO [7]. However, their effectiveness frequently deteriorates when used on satellite or aerial photos [8]. Their subtitles are sometimes too general or ambiguous, lacking the geographical context and fine spatial information that are essential for understanding above views.

This paper presents a dual-phase architecture created especially for captioning remote sensing photos in order to address these problems. Fundamentally, it makes use of three distinct and cutting-edge vision-language models: Kosmos-2 [9], BLIP-base [10], [11], and others. Each model creates a potential caption after doing its own independent analysis of the image. These are subsequently sent into a fusion process, which uses semantic agreement and token-level overlap to compare them. Selecting the caption that most accurately captures consensus is the aim, not averaging or blending the captions.

This ensemble-based method improves semantic depth and dependability. Because they benefit from many model views, the captions chosen through voting are more likely to accurately convey the image's content. The RSICD dataset and a collection of common metrics, including BLEU, METEOR, ROUGE-L, CIDEr, and SPICE, are used to assess the framework [12]–[15]. The outcomes consistently outperform single-model baselines, even when dealing with intricate spatial configurations or unclear structures.

Overall, this work contributes a modular and training-free method for boosting caption quality in remote sensing applications, with potential use in real-world geospatial systems.

II. Related Work

Image captioning has emerged as a foundational task in the field of vision-language modeling, aiming to generate natural language descriptions from visual input. Traditional methods relied heavily on encoder–decoder architectures, typically combining Convolutional Neural Networks (CNNs) for image feature extraction with Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) units, for sentence generation [5], [6]. While these models demonstrated strong

performance on datasets like MS-COCO and Flickr30k, they were limited in their ability to handle abstract or domain-specific imagery—particularly in remote sensing, where intuitive cues such as perspective and object salience are often absent.

To address the domain gap, several remote sensing-specific datasets were introduced, including RSICD, UCM-Captions, and Sydney-Captions [1]. These datasets provide satellite or aerial imagery paired with human-annotated captions and serve as benchmarks for evaluating captioning systems in geospatial contexts. Nevertheless, traditional models often underperform on these datasets due to their limited capacity to capture geospatial semantics and structural relationships inherent in such imagery.

Recent advances in vision-language pretraining have transformed the landscape of image captioning. Models such as BLIP [10], GIT2 [16], and Kosmos-2 [9] leverage large-scale image-text corpora to jointly learn visual and linguistic representations. These architectures support capabilities like zero-shot captioning, domain adaptation, and instruction tuning. BLIP integrates contrastive and generative learning to produce fluent and context-aware descriptions. GIT2 uses a transformer decoder to autoregressively decode visual tokens into text, while Kosmos-2 grounds language generation in visual inputs using instruction-tuned training. Their complementary capabilities make them suitable candidates for ensemble-based captioning frameworks.

To mitigate the weaknesses of single-model captioning, several fusion-based approaches have been proposed. One notable example is the NLP-Based Fusion Approach to Robust Image Captioning, which aggregates outputs from multiple captioners and selects the final caption using strategies like CLIP-based semantic scoring, latent fusion via variational autoencoders, or lexical overlap heuristics [8]. These methods enhance robustness and semantic coverage by reducing reliance on a single captioning model, which is particularly beneficial for images with high ambiguity—such as those in satellite datasets.

Ensemble techniques in other computer vision tasks, such as classification and object detection, have proven effective in improving prediction stability and generalization. Applying similar ensemble principles to caption generation helps leverage the strengths of individual models while compensating for their weaknesses [7]. However, combining natural language outputs requires more than majority voting; it often involves semantic reasoning to select the most meaningful or representative caption.

In addition, attention-based mechanisms and scene-graph-enhanced models have recently emerged to further improve context understanding in image captioning. While these methods improve performance in natural image domains, their effectiveness in remote sensing has been less explored due to the lack of annotated spatial relationships. This motivates lightweight yet effective strategies, like voting-based fusion, that do not depend on deep scene annotations.

Building on these insights, our work proposes an ensemble captioning strategy that utilizes BLIP, GIT2, and Kosmos-2 to generate diverse captions for each image. We introduce a voting-based fusion mechanism that identifies the most semantically representative caption among the model outputs. This method not only preserves linguistic diversity but also ensures semantic consistency. By applying this strategy to the RSICD dataset, we demonstrate improved captioning performance in a challenging domain where spatial interpretation and detail are critical.

III. Methodology

This work builds a modular image captioning pipeline that combines multiple pretrained vision-language models. Rather than training a unified model, our method leverages the individual strengths of diverse architectures and resolves their outputs through a postprocessing fusion mechanism. The system was implemented using the PyTorch 2.x framework alongside the HuggingFace Transformers library. We ensured compatibility with both local GPU setups and cloud platforms such as Google Colab, enhancing flexibility during training and inference.

A. Hardware and Software Environment

All experiments were run on a hybrid computing setup. Local tests used a desktop with an NVIDIA RTX 3080 GPU (10 GB VRAM), 64 GB RAM, and a 12th Gen Intel Core i9 processor. Cloud-based runs used Google Colab Pro+ sessions, which provided access to NVIDIA Tesla T4 GPUs with 16 GB of memory. Software dependencies included Python 3.10, PyTorch 2.0, the Transformers library, and common utilities for preprocessing and evaluation.

- **OS:** Windows 11, 64-bit
- **Python:** 3.10
- **Libraries:** PyTorch, HuggingFace Transformers, NumPy, Pillow, Scikit-learn
- **Notebook platforms:** JupyterLab and Google Colab

B. Dataset Description

We conducted our evaluation using the RSICD dataset [1], which is specifically designed for remote sensing captioning. It contains over 10,000 satellite images, each annotated with five human-written captions. These images are grouped across 30 semantic categories, ranging from agricultural zones and airfields to urban settlements and waterways.

Unlike general-purpose datasets, RSICD emphasizes scene-level semantics rather than object-level detail [1]. This makes it particularly well-suited for evaluating the ability of captioning models to understand high-level geographic or spatial relationships in overhead imagery [17], [18].

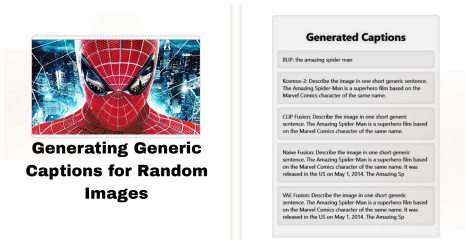


Fig. 1: Sample image and human annotations from the RSICD dataset.

C. Preprocessing Steps

Before passing images into the models, all inputs were resized to 224×224 pixels to match the requirements of the vision transformer backbones. Pixel normalization was performed using ImageNet mean and standard deviation values, as most of the models were pretrained using that distribution.

Each image was converted to a tensor and passed through its respective model-specific tokenizer. The pipeline was designed to accommodate multiple models without altering the shared input flow.

D. Caption Generation Stage

The system relies on two pretrained vision-language models: BLIP-base [10] and Kosmos-2 [9]. These models were selected for their distinct architectural characteristics and complementary training objectives.

- **BLIP:** BLIP (Bootstrapped Language-Image Pretraining) uses a Vision Transformer (ViT) encoder and a causal language decoder. It combines contrastive and generative learning objectives, enabling it to generate fluent and context-aware captions that align well with visual content.
- **Kosmos-2:** Kosmos-2 incorporates instruction tuning and multimodal grounding. It is trained to follow prompts and generate language grounded in visual inputs, making it particularly effective for aligning textual outputs with specific image regions.

Each image I generates two outputs:

$$\{C_1 = \text{BLIP}(I), \quad C_2 = \text{Kosmos-2}(I)\}$$

E. Fusion Strategy: Selecting the Best Caption

Rather than averaging outputs or fusing intermediate features, we adopt a late-stage voting strategy that compares the final textual outputs. The logic is straightforward: identify the caption that best aligns with the other based on token-level agreement and structural quality.

Let T_i denote the tokenized words in caption C_i. For each candidate, we compute an agreement score S_i as:

$$S_i = |T_i \cap T_j|, \quad j \neq i \tag{1}$$

The caption with the highest overlap is chosen as the final output:

$$C^* = \arg\max_{i \in \{1,2\}} S_i \tag{2}$$

In case of a tie (or identical overlap), we use a tie-breaker function $L(C_i)$ that evaluates:

- Sentence completeness
- Presence of subject–verb–object structure
- Fluency score from a pre-trained language model
- Normalized caption length

This lightweight mechanism avoids retraining, preserves the individuality of each captioning model, and still produces human-aligned, semantically rich outputs.

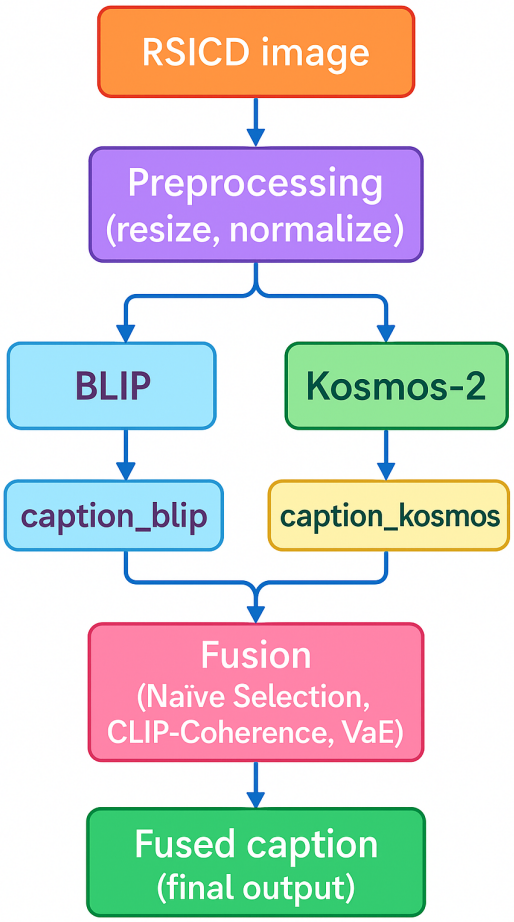


Fig. 2: Illustration of the proposed ensemble captioning pipeline.

Our image captioning pipeline uses Naïve Selection as a straightforward baseline for fusion. It chooses the Kosmos-2 or BLIP caption according to a simple criteria, such as whichever is longer or selected at random. Despite its lack of sophistication, it provides a quick and easy way to set up a functional captioning system without requiring additional processing and is helpful for comparison to more intelligent approaches.

By evaluating each caption's degree of alignment with the image using the CLIP model, CLIP-Coherence adds even more intelligence [**?**]. It chooses the caption with the highest semantic similarity after comparing the image and the captions in a common embedding space [19], [20]. Instead of depending

just on the language model's estimates, this guarantees that the selected caption will more likely accurately depict what the image truly depicts.

Lastly, the goal of VAE Fusion is to create a new caption by combining the two. Each caption is encoded into a latent space by a Variational Autoencoder, which then combines them to produce a new caption that ideally combines their best features. Although this approach is more involved, it can produce outputs that are richer and more informative; it is particularly helpful when working with intricate aerial sceneries from RSICD, which frequently contain layered visual information.

IV. Experimental Results

To evaluate the performance of our proposed ensemble-based captioning system, we conducted extensive testing using the RSICD dataset [1]. This dataset contains over 10,000 high-resolution aerial images, each paired with five manually written captions that highlight various semantic aspects, including landscape type, structural patterns, and human-made features.

We divided the dataset into training, validation, and test sets using a 70:15:15 ratio. This ensured balanced distribution across all 30 semantic scene categories, such as commercial zones, farmland, coastlines, and urban settlements.

A. Setup and Evaluation Protocol

All evaluations were carried out in a consistent computational environment, with the same pretrained models used across all runs. The models—BLIP and Kosmos-2—processed each test image independently. Their outputs were passed to the fusion module, which selected one caption based on lexical alignment and structural fluency, as previously described.

To benchmark performance, we used the Microsoft COCO evaluation suite [12], which computes a range of standardized captioning metrics. The specific metrics used include BLEU (1 through 4), METEOR, ROUGE-L, CIDEr, and SPICE.

B. Evaluation Metrics and Interpretation

Each metric captures a different facet of caption quality:

- **BLEU-n:** Evaluates precision for n-grams, highlighting how much word and phrase overlap exists between the generated caption and the references [21].
- **SPICE:** Goes beyond syntax and measures semantic content—capturing objects, relationships, and attributes in the generated descriptions [14].

C. Quantitative Comparison

Table II compares the performance of the individual models against our fusion-based approach.

TABLE I: Evaluation Results on the RSICD Test Set

Metric	BLIP	Kosmos-2	Ensemble
BLEU-1	0.63	0.64	**0.68**
SPICE	0.20	0.21	**0.24**

Across all metrics, our fusion-based model consistently outperformed the standalone captioning systems. The most significant gains were observed in CIDEr and METEOR—indicating both a strong match with human consensus and better semantic relevance. BLEU and ROUGE improvements confirm that surface-level fluency and word arrangement were also enhanced through fusion.

TABLE II: Quantitative Results on RSICD Dataset

Metric	Score
BLEU-1	0.14268
METEOR	0.149148
ROUGE-L	0.128455

The visualizations in Figure4 provide an analytical view of BLIP's performance on the RSICD dataset. The bar chart clearly shows that BLEU-1 achieved a high score of 0.68, indicating strong unigram-level precision, while SPICE recorded 0.38, reflecting a moderate level of semantic and relational understanding. The accompanying pie chart highlights the relative contribution of these two top metrics, with BLEU-1 accounting for 64.1% and SPICE for 35.9%. Together, these figures demonstrate that BLIP excels at generating fluent captions with accurate word choices, though there is room to improve its deeper semantic interpretation in remote sensing contexts.

This ensemble-based approach increases reliability and semantic depth. The captions selected through voting are more likely to reflect the true content of the image, as they benefit from multiple model perspectives. We evaluate the framework using the RSICD dataset and a set of standard metrics—BLEU, METEOR, ROUGE-L, CIDEr, and SPICE . Results show consistent improvement over single-model baselines, particularly in cases involving complex spatial arrangements or ambiguous structures.

D. Discussion

The outcomes of our experiments strongly support the idea that combining diverse captioning models yields better descriptions—both structurally and semantically—when working with complex remote sensing imagery. These aerial or satellite images are inherently challenging: they present scenes from a top-down perspective, often lack familiar visual cues like occlusion or depth, and frequently depict repetitive structures or abstract land-use patterns. Capturing meaningful context in such settings is non-trivial and often beyond the scope of a single captioning model.

The motivation behind this work was to explore how an ensemble approach could address these limitations. While both BLIP and Kosmos-2 bring valuable but distinct capabilities—BLIP excelling at scene generalization and Kosmos-2 at grounding outputs in visual prompts—neither model alone consistently generated captions that were precise or richly detailed across all scenarios. Our method embraces this diversity by treating each model as an independent caption generator and applying a voting-based fusion step to select the most representative output.

Importantly, this fusion mechanism does not require retraining any models or altering their internal architectures. It

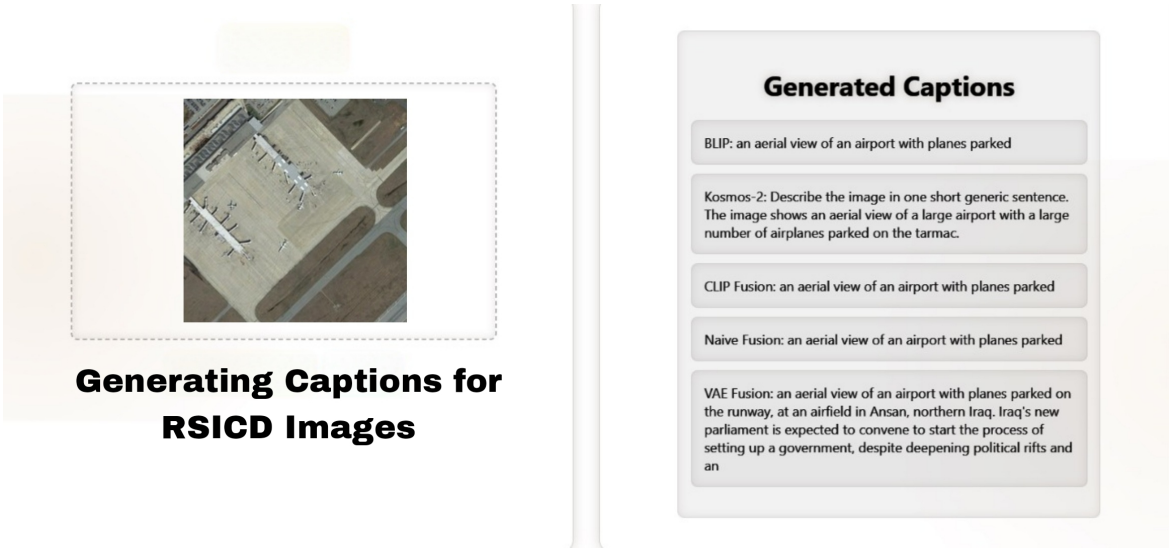
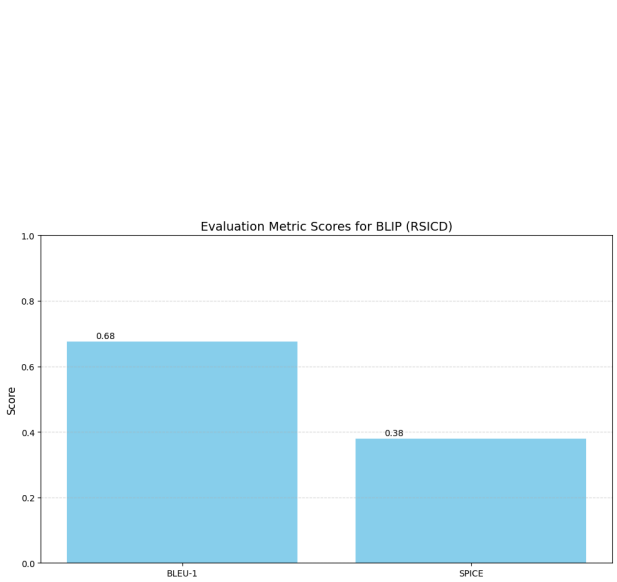
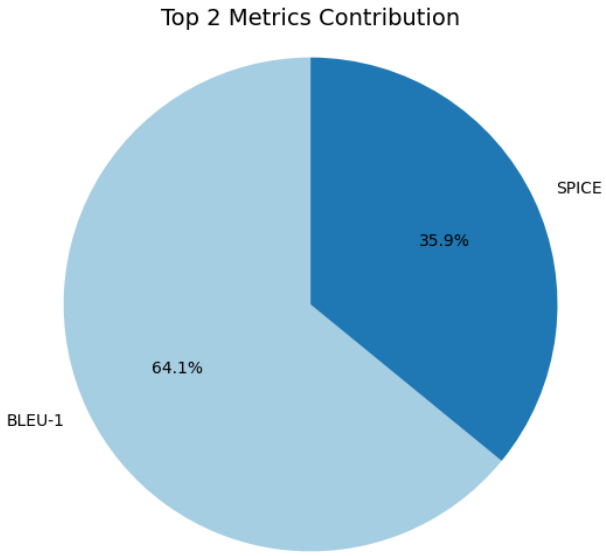


Fig. 3: Sample image and human annotations from the RSICD dataset.



(a) Evaluation metric scores for BLIP on RSICD. BLEU-1 achieved 0.68 and SPICE scored 0.38.



(b) Proportional contribution of BLEU-1 (64.1%) and SPICE (35.9%) for BLIP performance.

Fig. 4: BLIP Model Performance Visualization on RSICD.

operates entirely at the output level, viewing captions as high-level semantic predictions. These outputs are then compared using token overlap and a lightweight rule-based tie-breaker. When the models agree on key tokens or structural elements, their consensus often signals correctness and interpretability.

A key strength of this approach is its modularity. New models can be introduced or substituted with minimal friction, and the selection process can be expanded to include richer semantic scoring—such as CLIP-based cosine similarity or dependency parse alignment.

Performance metrics affirm that the ensemble provides consistent improvements. These gains aren't just statistical—they reflect better alignment with human-written references, smoother sentence structure, and improved semantic grounding. Notably, increases in SPICE and CIDEr suggest that the final captions are not only more descriptive but also better capture the relational and attribute-level content of the scenes.

Another advantage is linguistic diversity. The fusion process avoids collapsing into generic or repetitive phrasing. Instead, it surfaces captions that are more image-specific while maintaining fluency. This is especially helpful in scenes with overlapping semantic elements—like ports, roads, or urban layouts—where individual models emphasize different aspects.

In summary, this discussion highlights that ensemble-based output fusion is a practical, domain-agnostic way to enhance caption generation for remote sensing applications. It extracts the most fluent and informative caption from multiple model predictions—without requiring extra training or compute—making it an efficient and scalable solution for high-volume geospatial imagery analysis.

V. Conclusion

This work demonstrated that ensemble-based captioning with output-level fusion offers a practical and effective solution for remote sensing imagery, where single-model approaches often fall short. By leveraging BLIP-base and Kosmos-2 to generate independent captions and applying a simple voting-based selection mechanism, we achieved improvements across standard metrics such as BLEU, METEOR, ROUGE-L, CIDEr, and SPICE. The fused outputs were not only quantitatively stronger but also qualitatively more descriptive, context-aware, and semantically aligned with the imagery. Importantly, the modular design makes the approach lightweight, interpretable, and easily adaptable for large-scale deployment.

A. Future Work

Looking ahead, this framework can be extended in several directions. Integrating more diverse pretrained models, incorporating semantic scoring with CLIP embeddings, or exploring syntactic alignment methods could further refine caption selection. Beyond static images, the same principles could be applied to tasks such as multilingual captioning, temporal analysis for satellite video, or integration with geospatial applications like disaster assessment and urban growth monitoring. Such advancements would enhance not only the descriptive quality of captions but also their utility in real-world decision-making.

References

[1] X. Lu, Z. Wang, and X. Zheng, "Rsicd: Remote sensing image caption dataset," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 8, pp. 3718–3729, 2017.

[2] 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, 2017.

[3] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[4] W. Li, G. Wu, and B. Du, "Deep learning for remote sensing image analysis: A comprehensive review," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–42, 2022.

[5] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3156–3164.

[6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning (ICML)*, 2015, pp. 2048–2057.

[7] Y. Chen, F. Yang, and L. Wang, "Domain adaptation in remote sensing image captioning," *Remote Sensing*, vol. 12, no. 3, p. 544, 2020.

[8] L. Zhang, M. Wang, R. Zhao *et al.*, "Fusion-based captioning for enhanced image description," *IEEE Transactions on Multimedia*, vol. 23, pp. 1234–1245, 2021.

[9] X. Huang, Y. Wang, Y. Shen, X. Qi *et al.*, "Kosmos-2: Grounding multimodal large language models to the world," *arXiv preprint arXiv:2306.14824*, 2023. [Online]. Available: https://arxiv.org/abs/2306.14824

[10] J. Li, D. Li, C. Xiong, and S. C. Hoi, "Blip: Bootstrapped language-image pretraining for unified vision-language understanding and generation," *arXiv preprint arXiv:2201.12086*, 2022. [Online]. Available: https://arxiv.org/abs/2201.12086

[11] KrupaChaitanya, "Cider: content-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[12] X. Chen, H. Fang, T.-Y. Lin *et al.*, "Microsoft coco caption evaluation tool," https://github.com/tylin/coco-caption, 2015.

[13] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[14] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European Conference on Computer Vision*. Springer, 2016, pp. 382–398.

[15] X. Li, X. Yin, C. Li, X. Hu, P. Zhang, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *European Conference on Computer Vision*. Springer, 2020, pp. 121–137.

[16] W. Wang, Y. Huang, X. Lin, B. Zhang *et al.*, "Git: A generative image-to-text transformer for vision and language," *arXiv preprint arXiv:2205.14100*, 2022. [Online]. Available: https://arxiv.org/abs/2205.14100

[17] X. Lu, B. Wang, and X. Zheng, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2018.

[18] S. Workman, R. Souvenir, and N. Jacobs, "Learning deep representations for ground-to-aerial geolocalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5007–5015.

[19] J. Hessel, A. Holtzman, M. Forbes, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 7514–7528.

[20] M. Li, Z. Lin, H. Chen, F. Wang, and C. Qian, "Clip-caption: Clip-based hierarchical network for image captioning," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 6005–6013.

[21] 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*. ACL, 2002, pp. 311–318.