Domain generalized remote sensing scene captioning via Wikipedia geography articles

Authors

Institute

Abstract. In this study, we explore the performance impact of incorporating country-level text-based geographical information into a large-scale vision language model, fine-tuned for the captioning of optical remote sensing images. It is hypothesized that a model trained with country-level textual geographical context along with visual scenes will enhance its captioning capabilities when confronted with images from previously unseen countries or even continents, coupled with their respective geographical context. A Large Language and Vision Assistant (LLaVA) model was fine-tuned using optical images from European countries and tested on images from the other continents to evaluate its generalization capabilities. We report results of experiments conducted across 175 countries via the newly published Skyscript dataset, and show that even superficial geographical information obtained from Wikipedia articles can mitigate the cross-country domain shift by several points in terms of ROUGE score.

Keywords: Scene captioning · Domain Adaptation · Remote Sensing · Open vocabulary classification

1 Introduction

Equipping data driven models with robustness against domain shift is a critical challenge. Without proper counter-measures in place, trained models often suffer from performance degradation when confronted with data distributions that deviate from that of their training set [20]. This problem is exacerbated in remote sensing tasks such as land cover monitoring and scene description, where the extreme visual diversity of global landscapes can lead to critically significant distribution shifts.

Several transfer learning strategies have been reported for tacking this issue; namely domain adaptation [22], domain generalization [5], to name a few, depending on the availability of labels or data from the target domain(s). Unsupervised domain adaptation in particular continues to enjoy the attention of the remote sensing community [7], where only the target domain data is available with no labels, and has been extensively explored across a large variety of applications (object detection, semantic segmentation, etc.) [2, 10, 17]. Domain generalization on the other hand assumes no availability of either labels or data from the target domain [12] and thus represents a more realistic scenario.

Moreover, owing to their immense success with computer vision applications [4], large-scale vision language foundation models have also made a strong entry into the field of remote sensing image analysis [3,6,14,15]. Specifically, the advent of multi-modal models like CLIP [18] has revolutionized the integration of textual and image embeddings. Several models have been subsequently developed for remote sensing tasks such as visual question answering; e.g. GeoChat [11], RSGPT [9], and RS-LLaVA [1].

This paper focuses on addressing the domain shift in the context of scene captioning via a large-scale vision-language model; however distinctly from the state-of-the-art, higher generalization capacity is pursued not through target domain data, or target (pseudo-)labels, but rather via country-level geographic metadata in the form of Wikipedia articles about the target sample's country geography. We hypothesize that this contextual information that is almost always accessible via the image acquisition coordinates, provides additional discriminative signals. We adapt our model architecture to a dual-input network that separately processes and then fuses visual and text based geographic information. By incorporating this metadata into our multi-modal model, we aim to enhance the model's awareness of geographic contexts, thereby improving its ability to generalize across different domains/countries. This method allows the model to leverage rich, context-specific information that is not directly present in the images themselves but is highly relevant to understanding and describing the scenes.

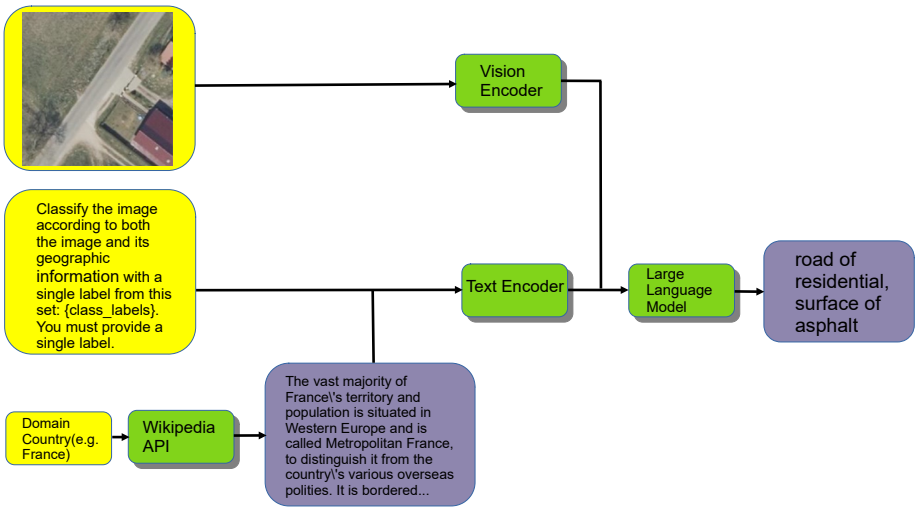

Fig. 1. Model architecture (yellow denotes inputs, green models or functions and purple represents outputs).

The investigated approach has been validated using the newly published Skyscript [21] dataset, comprising images from 175 countries. The results that have been obtained underline the potential benefits of incorporating country-level geographic description into multi-modal classification models to mitigate the negative effects of domain shift.

2 Proposed Approach

The model we have used in this study is LLaVA [16], an open-source multi-modal model capable of understanding simultaneously visual and textual inputs. In essence, LLaVA comprises two key components: a vision encoder and a language model called Vicuna, which is an extension of the Llama 2 architecture. The vision encoder is built upon a pre-trained CLIP ViT-L/14 model, which excels in understanding visual content. Meanwhile, Vicuna serves as the language processing backbone, adept at handling textual data.

During operation, the vision encoder processes images and converts them into embedding vectors, leveraging its pre-trained knowledge to extract meaningful visual features. Simultaneously, the language model Vicuna processes textual inputs, generating corresponding embedding vectors. Notably, these embedding vectors share the same dimensional space, facilitating seamless integration of visual and textual information within the model. This integration enables LLaVA to effectively understand and interpret multi-modal inputs, bridging the gap between visual and linguistic understanding. Even though the base model is highly effective, it still requires adaptation to specific tasks through additional training. However, due to its size (with billions of parameters),

we resorted to fine-tuning, specifically through LoRA [8] (Low-Rank Adaptation) with the end goal of scene captioning. Instead of adjusting all the parameters, LoRA introduces low-rank matrices that represent a simpler, smaller part of the model. During fine-tuning, only these low-rank matrices are updated, thus significantly reducing the amount of computational resources and data required. This approach renders the fine-tuning process faster and more cost-effective, enabling the model to be adapted to new tasks with minimal overhead while maintaining high performance.

Building upon this foundation, we introduced a novel element to the scene captioning process in the form of geographic information. In particular, we augmented our dataset with the country-level geographic metadata associated with each image, hypothesizing that such contextual information could provide additional discriminative signals to our model. The geographic information of each country was extracted through the publicly available API from its respective Wikipedia article (i.e. geography section) and was added to the model's prompt.

We designed a dual-input network that processes visual and geographic information separately, before fusing them for the final classification decision. This design allows the model to learn not only from the appearance of the scenes but also from the geographical context, potentially capturing the unique characteristics of scenes that are country-specific.

3 Experiments

The main objective of the experiments is to reveal the positive effects of geographic metadata in the scene captioning on remote sensing images for different domains. Two runs were realized, one with a baseline LLaVA model with no geographic metadata and a model with geographic metadata. The metric used to evaluate the models is the ROUGE-Recall score [13], which calculates performance by measuring the overlap between the predicted and reference summaries, specifically focusing on the recall aspect. We have given a 12 token limit for the model to generate so that the ROUGE-Recall score would not give inaccurate results. This means it assesses how many of the important units (such as n-grams, words, or sequences) in the reference summaries are successfully captured in the generated summaries, providing an indication of the model's ability to include relevant information.

3.1 Dataset

SkyScript [21] is a comprehensive vision-language dataset specifically designed for remote sensing images. As it encompasses 175 countries, it exhibits a significant amount of domain shift across countries and continents. It was constructed to address the absence of a large-scale, semantically diverse image-text dataset required for developing VLMs for remote sensing images. Unlike natural images, remote sensing images and their associated text descriptions cannot be efficiently collected from the public Internet at scale. The dataset is constructed by using geo-coordinates to automatically connect open, unlabeled remote sensing images with the rich semantic information available in OpenStreetMap.

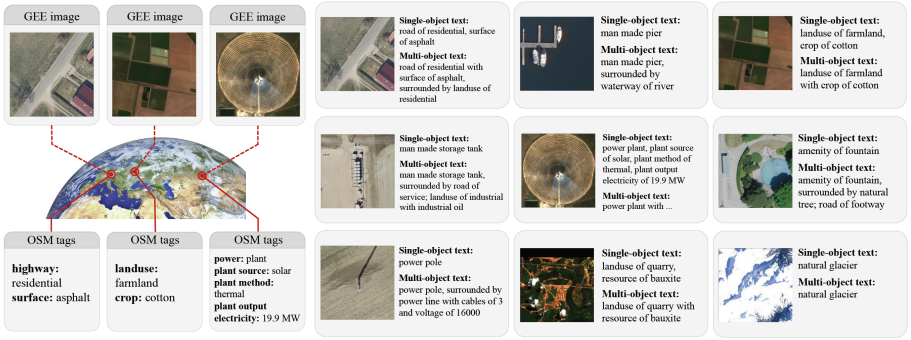

Fig. 2. Skyscript sample examples [21]

SkyScript comprises of 2.6 million image-text pairs and covers 29,000 distinct semantic tags from 175 countries. SkyScript, forming a multi-source, multi-resolution image pool with ground sampling distance (GSD) ranging from 0.1

m/pixel to 30 m/pixel of various sizes. In the present study, only a small subset of the dataset was extracted for scene captioning, since there is a significant number of unique captions in the dataset. Only the captions that have a frequency over 15,000 were used. The number 15,000 was determined empirically, because lower numbers had an excessive number of classes.

The model was trained on the task of scene captioning with 37 unique captions (Table 1).

Table 1. The captions used during the experiments.

leisure land of pitch	landuse of retail	landuse of residential
road of turning loop	natural water	landuse of commercial
airport of taxiway	highway of freeway junction	tunnel of culvert
landuse of meadow	man made pier	natural wetland
road of stop	golf hole	railway of level crossing
power tower	road of turning circle	road of service
power switch	leisure land of park	road of crossing
landuse of farmland	waterway of canal	road of residential
waterway of river	building of residential	power pole
place of village	natural peak	place of hamlet
landuse of farmyard	building of house	natural tree
building of farm	building of barn	amenity of parking space
amenity of school		

Training test consists of images from Germany, France, Switzerland, Spain, Finland. And the testing set consists of the countries all over the world.

3.2 Settings

There are two different training settings we have built to compare the performances:

- LLaVA
- LLaVA with geography metadata

Hyperparameters chosen for these settings can be seen in Table 2.

Hyperparameter	**LLaVA**	**LLaVA with Geo**
Learning Rate	2e-3	2e-4
Batch Size	8	8
Learning Rate Scheduler	Cosine	Cosine
Optimizer	Adam	Adam

Table 2. Hyperparameters for training both LLaVA and LLaVA with Geo.

Since this is a scene captioning problem, the prompts given to the LLaVA model has been modified for this downstreaming task. The template of the prompt is taken from [19]. And finally for evaluation only the predictions with ROUGE-Recall score ≥ 0.75 has been deemed as accurate and predictions with ROUGE-Recall score ≥ 0.5 have been deemed to be half-accurate. The total score is calculated by Eq. (1).

$$
\begin{aligned}
R &= \text{ROUGE-Recall score} \\
A &= \text{Number of accurate guesses } (R \geq 0.75) \\
H &= \text{Number of half-accurate guesses } (0.5 \leq R < 0.75) \\
W &= \text{Number of wrong guesses } (R < 0.5)
\end{aligned}
$$

The total score S is calculated as:

$$S = \frac{A + 0.5H}{A + H + W} \times 100 \tag{1}$$

Establishing a baseline The text input consists only of the classification question and the possible answers. The training was conducted on a single A100 GPU for a week in both settings. The prompt of the setting can be seen in Figure 3

Fig. 3. Prompt of the first setting

<image> \n Classify the image with a single label from this set: class_labels. You must provide a single label.

The scores calculated via Eq. (1) for this type of training are provided in Fig. 4

LLaVA with geography metadata Text input from the second setting is constructed by combining both geographic metadata and the classification question.The prompt of the setting can be seen in Figure 5

The scores for the model trained with geographic metadata calculated via Eq. (1) for this type of training are shown in Fig. 6

3.3 Results & Discussion

At Fig. 7 you can see a map that shows the difference in performance of geographical information data entered to the multimodal vs not. The positive values

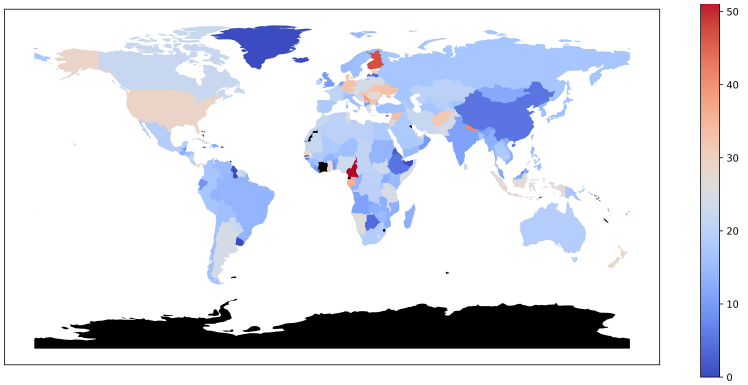

Fig. 4. Baseline results on each country

Fig. 5. Prompt of the second setting

<image>\n<geo_info>...</geo_info>Classify the image according to both the image and its geographic information with a single label from this set: class_labels. You must provide a single label.

indicate the performance increase with the geographical information and negative indicates the opposite.

It can be seen that except for a few countries, geographical information has a positive impact on performance. All the results for both settings are provided in Table 3. With the addition of geographical metadata in terms of all the countries the score increased over 47%. There can be various reasons behind the negatively affected countries, their Wikipedia page might have an insufficient and inaccurate geographical information or the country might have regions so diverted from each other (to solve this problem the geography of countries can be switched with geography of regions).

4 Conclusion

We have proposed an approach to tackle domain adaptation using multimodal models. Giving multimodal model geographic metadata of the image boost the performance and reduces the harm of the domain shift in remote sensing image captioning. As far as we know, this is the first implementation of Wikipedia geography data to reduce the geographical domain shift. In future work, this implementation can be expanded to other computer vision tasks such as semantic segmentation, object detection and visual question answering or the same setting can be done on different multimodal models other than LLaVA to see if they behave the same.

Table 3. ROUGE scores for both the baseline captioning model and for the proposed approach reinforced with Wikipedia geography articles.

Country	Baseline	Ours	Country	Baseline	Ours	Country	Baseline	Ours	Country	Baseline	Ours	Country	Baseline	Ours
Africa														
Algeria	20	29	Angola	11	27	Benin	12	18	Botswana	4	21	Burkina Faso	14	21
Burundi	16	26	Cameroon	52	39	C. Afr. Rep.	21	26	Chad	21	27	Congo	14	21
Dem. Rep. Congo	20	29	Djibouti	16	0	Egypt	18	22	Eritrea	2	15	Ethiopia	5	21
Gabon	36	44	Ghana	27	32	Guinea	16	33	Guinea-Bissau	7	28	Kenya	15	22
Lesotho	26	40	Liberia	9	19	Libya	20	26	Madagascar	13	24	Malawi	25	35
Mali	18	29	Mauritania	22	55	Morocco	22	3	Mozambique	13	23	Namibia	26	30
Niger	16	27	Nigeria	23	25	Rwanda	18	18	Senegal	27	39	Sierra Leone	14	23
Somalia	25	1	Somaliland	0	0	South Africa	19	18	S. Sudan	23	34	Sudan	14	29
Tanzania	24	33	Togo	8	25	Tunisia	25	37	Uganda	13	20	Zambia	13	32
Zimbabwe	15	26												
America														
Argentina	24	34	Belize	7	7	Bolivia	16	23	Brazil	14	23	Canada	22	38
Chile	16	22	Colombia	17	29	Costa Rica	18	30	Cuba	27	37	Dominican Republic	11	22
Ecuador	9	22	El Salvador	25	50	Guatemala	13	21	Guyana	0	25	Haiti	14	25
Honduras	16	32	Jamaica	0	7	Mexico	19	29	Nicaragua	22	32	Panama	30	30
Paraguay	22	32	Peru	17	29	Suriname	23	23	United States	29	45	Uruguay	0	12
Venezuela	15	0												
Asia														
Afghanistan	32	40	Armenia	22	25	Azerbaijan	17	30	Bangladesh	17	25	Bhutan	7	23
Brunei	0	33	Cambodia	19	34	China	5	19	Georgia	19	40	India	11	20
Indonesia	27	35	Iran	22	34	Iraq	18	29	Israel	27	41	Japan	16	27
Jordan	27	36	Kazakhstan	20	26	Kyrgyzstan	17	26	Laos	17	28	Lebanon	31	36
Malaysia	12	20	Mongolia	12	20	Myanmar	13	25	N. Korea	15	29	Nepal	41	46
Oman	10	25	Pakistan	24	33	Palestine	37	37	Philippines	15	26	Qatar	33	66
S. Arabia	18	24	S. Korea	16	31	Sri Lanka	22	30	Syria	34	43	Taiwan	17	31
Tajikistan	21	31	Thailand	18	30	Turkmenistan	17	22	Turkey	16	24	U.A.E.	23	35
Uzbekistan	19	26	Vietnam	13	23	Yemen	15	22						
Europe														
Albania	12	32	Armenia	22	25	Austria	26	55	Belarus	24	35	Belgium	19	34
Bosnia and Herz.	26	30	Bulgaria	33	42	Croatia	21	33	Czechia	24	39	Denmark	33	33
Estonia	30	41	Finland	47	51	France	21	27	Germany	33	52	Greece	25	39
Hungary	32	37	Iceland	0	16	Ireland	19	33	Italy	16	31	Kosovo	28	35
Latvia	4	11	Lithuania	24	33	Luxembourg	28	40	Moldova	25	25	Montenegro	23	37
Netherlands	11	21	North Macedonia	10	20	Norway	15	25	Poland	24	31	Portugal	19	29
Romania	28	32	Russia	17	25	Serbia	38	44	Slovakia	23	36	Slovenia	24	30
Spain	18	51	Sweden	16	24	Switzerland	17	51	Ukraine	33	43	United Kingdom	10	16

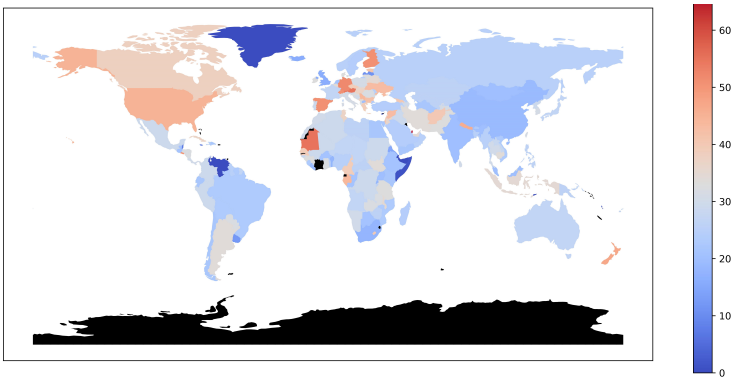

Fig. 6. LLaVA with geo metadata on each country

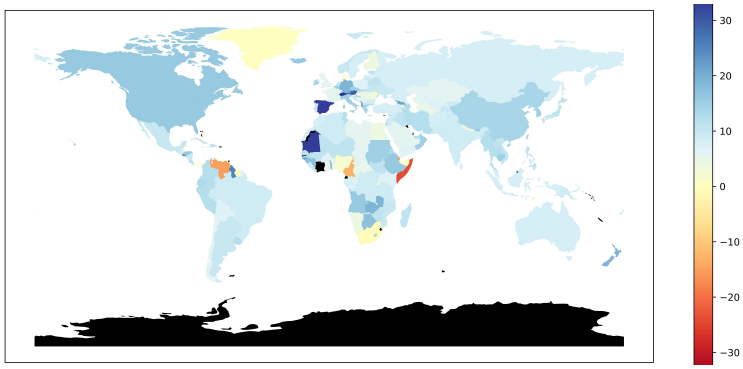

Fig. 7. Differences of geo scores and baseline scores by country

References

1. Bazi, Y., Bashmal, L., Al Rahhal, M.M., Ricci, R., Melgani, F.: Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery. Remote Sensing **16**(9), 1477 (2024)
2. Chen, J., He, P., Zhu, J., Guo, Y., Sun, G., Deng, M., Li, H.: Memory-contrastive unsupervised domain adaptation for building extraction of high-resolution remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–15 (2023)
3. Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., Shi, Z.: Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. IEEE Transactions on Geoscience and Remote Sensing (2024)
4. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems **36** (2024)
5. Durakli, E., Aptoula, E., Bosilj, P., Fox, C.: A domain generalized mask r-cnn for building instance segmentation. In: Proceedings of IEEE IGARSS. Athens, Greece (2024)

6. Fang, L., Kuang, Y., Liu, Q., Yang, Y., Yue, J.: Rethinking remote sensing pre-trained model: Instance-aware visual prompting for remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–13 (2023)
7. Hong, D., Zhang, B., Li, H., Li, Y., Yao, J., Li, C., Werner, M., Chanussot, J., Zipf, A., Zhu, X.X.: Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. Remote Sensing of Environment **299**, 113856 (2023)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
9. Hu, Y., Yuan, J., Wen, C., Lu, X., Li, X.: Rsgpt: A remote sensing vision language model and benchmark. arXiv preprint arXiv:2307.15266 (2023)
10. Ismael, S.F., Kayabol, K., Aptoula, E.: Unsupervised domain adaptation for the semantic segmentation of remote sensing images via one-shot image-to-image translation. IEEE Geoscience and Remote Sensing Letters (2023)
11. Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., Khan, F.S.: Geochat: Grounded large vision-language model for remote sensing. arXiv preprint arXiv:2311.15826 (2023)
12. Liang, C., Li, W., Dong, Y., Fu, W.: Single domain generalization method for remote sensing image segmentation via category consistency on domain randomization. IEEE Transactions on Geoscience and Remote Sensing (2024)
13. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
14. Liu, C., Zhao, R., Chen, J., Qi, Z., Zou, Z., Shi, Z.: A decoupling paradigm with prompt learning for remote sensing image change captioning. IEEE Transactions on Geoscience and Remote Sensing (2023)
15. Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J.: Remoteclip: A vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing (2024)
16. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
17. Liu, W., Liu, J., Luo, B.: Unsupervised domain adaptation for remote sensing vehicle detection using domain-specific channel recalibration. IEEE Geoscience and Remote Sensing Letters (2023)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
19. Roberts, J., Lüddecke, T., Sheikh, R., Han, K., Albanie, S.: Charting new territories: Exploring the geographic and geospatial capabilities of multimodal llms. arXiv preprint arXiv:2311.14656 (2023)
20. Tuia, D., Persello, C., Bruzzone, L.: Domain adaptation for the classification of remote sensing data: An overview of recent advances. IEEE geoscience and remote sensing magazine **4**(2), 41–57 (2016)
21. Wang, Z., Prabha, R., Huang, T., Wu, J., Rajagopal, R.: Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5805–5813 (2024)
22. Xu, Q., Shi, Y., Yuan, X., Zhu, X.X.: Universal domain adaptation for remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–15 (2023)