International Journal of Informatics, Information System and Computer Engineering

Experimental Evaluation of CLIP-Based Zero-Shot Classification of Imbalanced Remote Sensing Scenes: Addressing Quantity Disparities in Data

Tanvir Ahmed, Tanha Asfika Jaman **, Shekh Ifteesham Iftee **, Tanjoy Mahmud *, Ekra MD Emadur Rahman *, Hossain MD Maruf **

* School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing, China
** School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing, China
*Corresponding Email: tanvirahmed@nuist.edu.cn

ABSTRACTS

This paper presents a zero-shot learning framework based on Contrastive Language Image Pretraining (CLIP) for Remote Sensing Scene Classification (RSSC). The proposed method addresses the challenge of imbalanced image quantities across different categories, which is often encountered in practical ap-plications. Traditional zero-shot learning methods in RSSC leverage pre-trained word embeddings to extract semantic features from category names or descriptions, which are then fixed during the learning process without adaptation to visual features. This leads to a gap between visual and semantic representations. We have integrated the Vision Transformer with CLIP to enhance the alignment between visual and semantic features. Extensive experiments conducted on WHU-RS19 dataset demonstrate the effectiveness of the proposed framework, show-casing improved classification performance and generalization capabilities.

ARTICLE INFO

Article History:
Received 30 Sept 2024
Revised 10 Oct 2024
Accepted 06 Nov 2024
Available online 17 Dec 2024
Publication date 01 June 2025

Keywords:
Remote Sensing Scene Classification (RSSC),
Contrastive Language Image Pretraining (CLIP),
Zero-Shot Learning (ZSL)

DOI: https://doi.org/10.34010/injiiscom.v6i1.14164
p-ISSN 2810-0670 e-ISSN 2775-5584

1. INTRODUCTION

Remote sensing scene classification (RSSC) is a critical task in Remote Sensing (RS) scene analysis, involving the categorization of aerial or satellite scenes into predefined semantic categories that represent various land cover types, urban structures, or natural phenomena. This classification process is essential for numerous applications, including environmental monitoring, urban planning, disaster management, and agricultural assessment (Chen and Tsou 2021; Zhu et al. 2023; Jin et al. 2022). The traditional approach to RSSC relies heavily on manually extracted features and conventional machine learning algorithms, which often require substantial labelled data and exhibit limited generalization capabilities.

Recent advancements in deep learning, have significantly improved the performance of RSSC tasks. CNNs are capable of automatically learning hierarchical features from raw image data, thereby reducing the need for manual feature extraction and achieving state-of-the-art results in various image classification problems (Li et al. 2021). Despite these advancements, with the rapid blossoming of RS technology, an ever-growing volume of data brings the bulk of costs in manpower and material resources. This poses new challenges in reducing the burden of manual labour and improving efficiency. For this reason, more intelligent and automatic approaches for diverse RS applications must be developed.

Inspired by the human ability to recognize new objects by relating them to known concepts, Zero-Shot Learning (ZSL) has emerged as a promising solution. ZSL seeks to classify instances from unseen classes by leveraging knowledge transferred from seen classes using semantic information like attributes or textual descriptions (Romera-Paredes and Torr 2017; Xian et al. 2016). In the context of RSSC, the first ZSL method utilizing word2vec embeddings to map both seen and unseen classes into a shared semantic space, allowing for the classification of unseen categories through label propagation on a semantic graph (Li et al. 2017). Recent advancements in ZSL for RSSC have focused on improving the alignment between visual and semantic spaces. For instance, a semi-supervised Sammon embedding algorithm has been used to enhance the synthesis capabilities of unseen class prototypes (Quan et al. 2018). A deep cross-modal embedding network with locality-preservation constraints has been introduced to address class structure inconsistency (Li et al. 2021). Additionally, a distance-constrained semantic autoencoder has been proposed to better align visual features with semantic representations, resulting in higher classification accuracy for unseen classes (Wang, Peng, and Baets 2021).

Nevertheless, existing methods predominantly rely on a pretrained word2vec and BERT model on the Wikipedia corpus to derive semantic embeddings from category names or descriptions. During the zero-shot

DOI: https://doi.org/10.34010/injiiscom.v6i1.14164
p-ISSN 2810-0670 e-ISSN 2775-5584

learning process, these semantic embeddings are preprocessed and remain static, without aligning to visual features. This approach can result in limited representational capability of the extracted semantic embeddings and significant discrepancies between visual and semantic features. Furthermore, previous methods often employ shallow networks for learning visual and semantic features, presenting an additional challenge.

In order to address these issues, we have introduced a Vision-Transformer based model for ZSL RS scene classification. Recently, the integration of vision-language models has further advanced ZSL in RSSC. Contrastive Language-Image Pre-Training (CLIP), which learns visual representations through language supervision, have demonstrated exceptional performance in zero-shot and few-shot learning tasks by aligning visual and textual features (Radford et al. 2021; Li et al. 2022a). Vision Transformers (ViTs), which approach image classification as a sequence prediction problem, have also shown superior performance by capturing long-range dependencies in images (Dosovitskiy et al. 2020). These advancements offer distinct advantages for RSSC. CLIP excels by using large-scale pre-trained models to align visual and textual data, reducing the need for extensive labelled datasets and handling various remote sensing tasks efficiently (Jia et al. 2021). However, the efficiency of CLIP in handling large datasets through pre-trained models provides an edge in terms of scalability and application in diverse scenarios.

Most of the current Zero-shot Remote Sensing Scene Classification (ZSRSSC) methods are validated on the RSSDIVC dataset, which has a balanced number of images for each category. Therefore, these methods did not consider the impact of imbalanced image quantity of categories. However, this problem often exists and significantly impacts performance in practical application scenarios. Hence, advancing the integration and enhancement of the implementation mechanisms for ZSRSSC could represent an additional avenue for the development.

The main contributions of this paper are:

1.We have Proposed a Zero-Shot Learning framework based on CLIP for Remote Sensing Scene Classification in this paper.

2.Then, We have conducted our experiments on a dataset where the quantity of the RS scene is not equal for all the classes addressing the issue of lack of methods based on imbalanced scene quantity, eradicating the negative impact for practical applications.

This paper mainly includes five sections. The next section provides the related works. And, then Section 3. Provides the methodology proposed in this paper. And, in Section 4. We have provided the experimental details, results and discussions. Finally, chapter 5 concludes the paper.

DOI: https://doi.org/10.34010/injiiscom.v6i1.14164
p-ISSN 2810-0670 e-ISSN 2775-5584

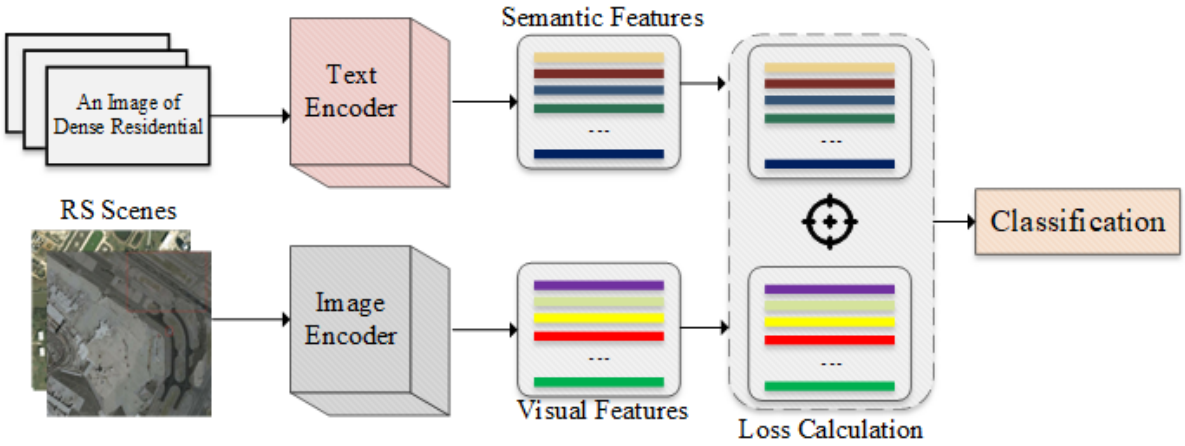

Fig. 1. CLIP Overview

2. RELATED WORKS

2.1. Remote Sensing Scene Classification

Remote Sensing Scene Classification (RSSC) is a critical task in interpreting high-resolution remote sensing imagery. With the rapid advancements in satellite remote sensing technology in recent years, RSSC has also seen considerable progress. Unlike conventional object classification tasks, remote sensing scenes represent highly abstract semantic concepts that do not necessarily align with specific land feature types. As a result, traditional pixel-level or object-level classification methods are challenging to apply directly to these scenes (Wang et al. 2024). To develop more discriminating features for describing remote sensing scenes, researchers have conducted extensive studies and achieved significant advancements. Notably, descriptors based on artificial design leverage prior knowledge, offering benefits such as low computational requirements and strong interpretability. For example, Zhu et al. introduced a knowledge-based land pattern description framework that employs an adaptive gradient perception mechanism and a land pattern cognitive model to capture the internal and external relationships between different land cover types (Zhu et al. 2022). Penatti et al. then evaluated the generalization capability of deep features (ConvNets) in two novel contexts: aerial and remote sensing images. They assessed the effectiveness of these deep learning features in classifying both aerial and remote sensing images (Penatti, Nogueira, and Santos 2015). Chaib et al. utilized a Visual Geometry Group network (VGGNet) as a feature extractor to select more representative deep features, thereby optimizing the representation of remote sensing scenes (Chaib et al. 2017). At the same time, although these methods have demonstrated good performance on specific datasets, they exhibit significant limitations in practical applications. These limitations are primarily due to the following: 1) The visual features of the same urban remote sensing scene can vary greatly across different countries or

DOI: https://doi.org/10.34010/injiiscom.v6i1.14164
p-ISSN 2810-0670 e-ISSN 2775-5584

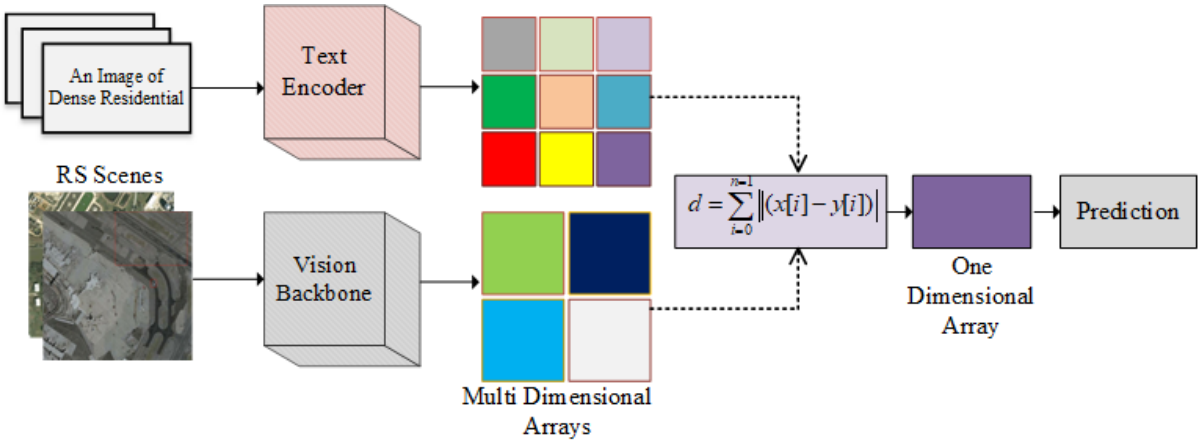

Fig. 2 Schematic Diagram of the Proposed Framework.

even different regions within the same country, influenced by factors such as eco-nomic development levels and cultural differences (Gawlikowski et al. 2022). Therefore, classification models trained on specific datasets are typically not transferable and cannot be directly applied to the classification of remote sensing scenes in different cities. Additionally, modern urban remote sensing scenes are diverse and continuously evolving, while existing models are limited to classifying only the labeled scene categories provided in the training set. These models lack the capability to classify other unlabeled or newly emerging scene categories, indicating poor generalization ability and scalability.

2.2. ZSL in Remote Sensing Scene Classification

By ZSL in RS scene classification, it refers to classifying unseen RS scene classes with the assistance of unseen semantic information by learning the correspondence between the seen scene class images and their semantic information. Li et al. introduced the first zero-shot learning based remote sensing scene classification method. They have leveraged a word2vec model pretrained on the Wikipedia corpus to extract semantic embeddings from category names. A semantic graph was then built to characterize the relationship between semantic classes (Li et al. 2017). Wang et al. proposed a distance-constrained semantic autoencoder to reduce the semantic gap between visual features, and semantic representations, which to some extent alleviates the do-main shift problem (Wang, Peng, and Baets 2021). Quan et al. designed a semi supervised Sammon embedding algorithm to transfer the unseen knowledge in the semantic space to the visual space, making it more consistent with the class structure in the visual space (Quan et al. 2018). Then Li et al. used generative adversarial networks (GANs) for zero-shot RS scene classification (Li et al. 2022b). Then, Ma et al. also proposed a ZSL framework for RS scene classification using generative adversarial networks (GANs) to better measure the reconstruction quality in Zero-shot RS scene classification (Ma et al. 2022). Despite significant advancements in the field, most of the

DOI: https://doi.org/10.34010/injiiscom.v6i1.14164
p-ISSN 2810-0670 e-ISSN 2775-5584

Table 1 Dataset Information

Dataset	Class Names
WHU-RS19	airport, bridge, river, forest, grassland, pond, parking lot, port, overpass, residential area, industrial area, commercial area, beach, desert, farmland, soccer field, mountain, park, and train station.

previous methods, only focused on either Global features or in local features exist in the RS scenes. Wang et al. proposed a zero-shot RS scene classification, fusing local and global features, introducing a weight mapping loss, quoting the necessities of combining local and global features in RS scene classification (Wang et al. 2024).

3. METHODOLOGY

In this section, we have presented the proposed method for zero-shot remote sensing scene classification based on vision-language models. First, we briefly discuss the traditional CLIP model and then we explain our techniques for ZSL RS scene classification. The CLIP model creates visual representations guided by language, as shown in Fig. 1. Given a set of N image-text pairs, the CLIP model aims to match the images with the correct text descriptions. To do this, the model uses a vision encoder to process images of remote sensing scenes and a language encoder to handle text descriptions, such as "an image of dense residential."

During training, the CLIP model generates visual and semantic features through their respective encoders. It then predicts a similarity matrix S, where each row represents the probability of matching one image with each of the N texts. The model is trained to increase the similarity scores. This is done by optimizing a cross-entropy loss over the similarity matrix.

For classification tasks with C categories, such as$(1, 2, \dots \dots, C)$, CLIP uses a prompt like "an image of a [CLASS]" to create text inputs T, where [CLASS] is replaced with the names or descriptions of each category. The text encoder generates semantic features for all classes, resulting in $F_t = E_t(T)$. For a batch of images III, the vision encoder produces visual features, resulting in $F_i = E_i(I)$. The classification probabilities are computed as follows:

$$P = Softmax(F_i + \frac{F_t^T}{\tau}) \quad (1)$$

Here, F_i and F_t are normalized, and their matrix product represents their cosine similarity. The parameter τ is a learnable scaling factor. Applying the $Softmax$ function yields a probability matrix P, where each row shows the likelihood of each image belonging to each class. The final classification is determined by choosing the class with the highest probability:

$$Y = argmax(P) \quad (2)$$

This method, illustrated in the Fig. 2, shows the entire process from handling remote sensing images and text descriptions through the encoders to the final classification output. We first start with the collection of images from remote sensing scenes. These images, which can

DOI: https://doi.org/10.34010/injiiscom.v6i1.14164
p-ISSN 2810-0670 e-ISSN 2775-5584

Fig. 3. RS Scene Samples

capture various environments such as dense residential areas, serve as the primary visual input for the analysis. Accompanying these images are textual descriptions or annotations that provide additional context and information about the scenes. The first major processing step involves the text encoder. The textual descriptions associated with the RS scenes are processed using a text encoder, which transforms the text into a multi-dimensional feature array. This transformation is crucial for enabling direct comparisons between text and image data. Here, T denote a textual description, and $E_t(T)$ represent the text encoder function. The text encoder outputs a feature vector $x \in R^N$:

$$X = E_t(T) \tag{3}$$

where N is the dimensionality of the feature space. This feature vector encapsulates the semantic information contained within the semantic information. Parallel to the text encoding process, the RS images are processed through the vision backbone. The vision backbone processes the scenes to extract meaningful features and represents them as a multi-dimensional array. Let I denote an input image, and $E_v(I)$denote the vision backbone function. The output feature vector $y \in R^M$is given by:

DOI: https://doi.org/10.34010/injiiscom.v6i1.14164
p-ISSN 2810-0670 e-ISSN 2775-5584

Table 2 OA (%), SD (%) and Kappa Coefficients for different Vision Backbones

Vision Backbones	OA (%) and SD (%)	Kappa Coefficient
ViT-B/32	67.36±3.4	0.6547
ViT-B/16	72.73±4.1	0.7123
ViT-L/14	75.65±3.8	0.6799

$$y = E_v(I) \quad (4)$$

here M is the dimensionality of the image feature space. This feature vector captures the visual characteristics of the RS scene. Then, comparing the multi-dimensional arrays from the text encoder and the vision backbone to measure their similarity. This comparison is quantified using a distance metric, which sums the absolute differences between corresponding elements of the feature vectors:

$$d = \sum_{i=1}^{min(N,M)} |x[i] - y[i]| \quad (5)$$

The distance metrics provide a measure of how similar or different the semantic and visual features are. The calculated distance metric d is then converted into a one-dimensional array z. This aggregation step simplifies the representation of the comparison results, enabling straightforward input into the prediction model. Finally, the one-dimensional array z is fed into a classifier layer. The final prediction $y^{\hat{y}}$is derived is as shown in Fig. 2.

4. EXPERIMENTAL SETTINGS & RESULTS ANALYSIS

4.1. Experimental Settings

The simulations are performed on a system with an Intel(R) Core (TM) i7-8750H CPU and a GeForce GTX1650 GPU with 16 GB of RAM. And the model is tested in WHU-RS19 dataset for three different vision backbones. The dataset consists of 19 categories of RS scenes, as described in Table 1 where each category consists of 50~61 RS scenes with dimension of 600×600 pixels, which indicates the imbalanced RS scene quantities per category. We have tested the framework under three vision backbones, ViT-B/32, ViT-B/16, and ViT-L/14.

4.2. Results Analysis

The performance of the proposed zero-shot remote sensing scene classification method using Vision Transformers (ViTs) is evaluated on the WHU-RS19 dataset. This dataset consists of 19 categories of high-resolution remote sensing scenes, as shown in Fig. 3, each containing between 50~61 RS scenes of 600×600 pixels. The performance is tested using three different vision backbones: ViT-B/16, ViT-B/32, and ViT-L/14. To quantitively evaluate the performance of the method, we have used several quantitative evaluation metrics including, the overall accuracy (OA) and standard deviation (SD), along with the Kappa coefficient

DOI: https://doi.org/10.34010/injiiscom.v6i1.14164
p-ISSN 2810-0670 e-ISSN 2775-5584

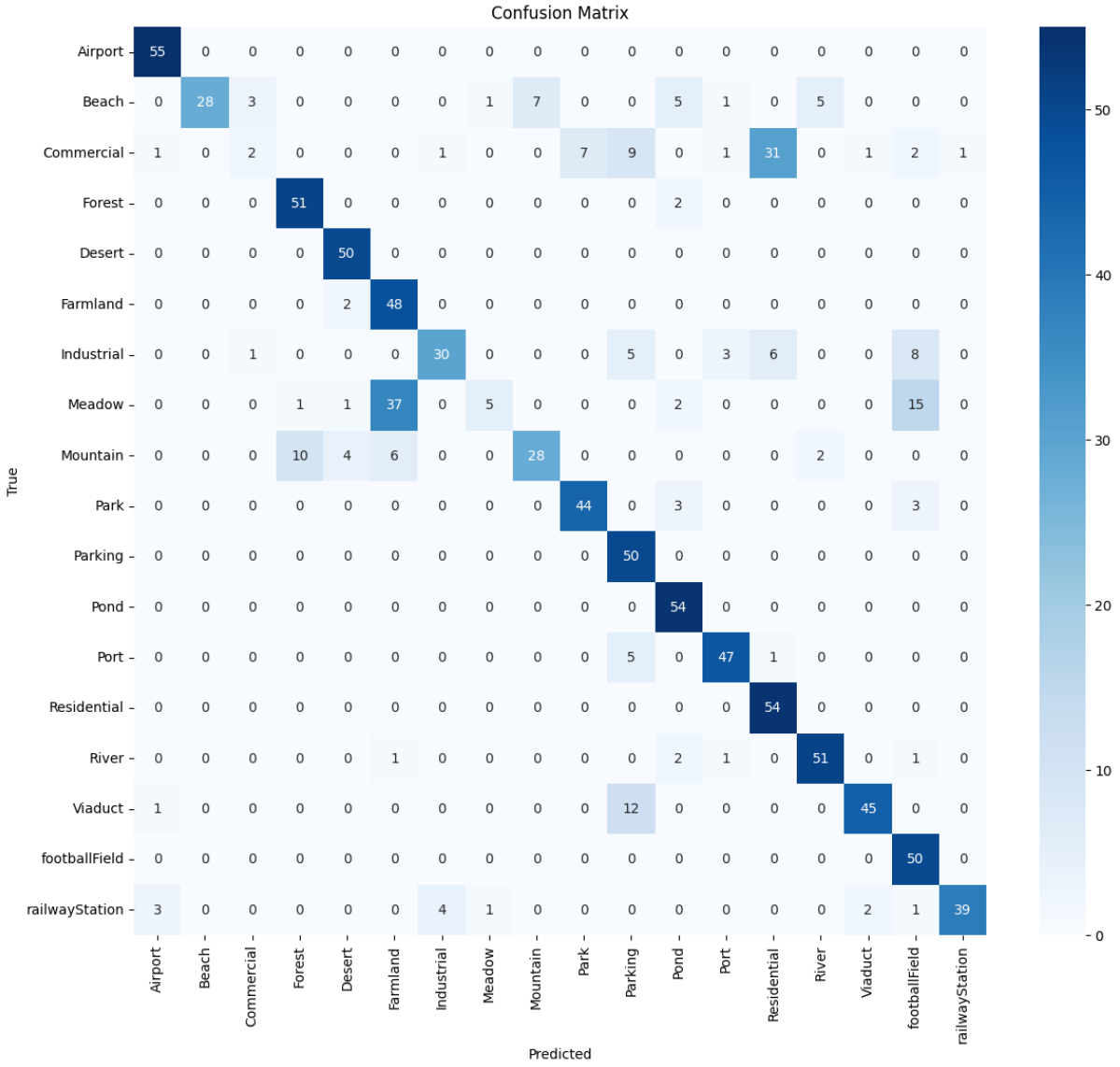

Fig. 4. Confusion Matrix of the Classification for ViT-B/16

and Confusion Matrix (CM) for each vision backbone. OA is a direct measure of the classification accuracy of the model on the entire dataset.

$$o_A = \frac{N_t}{N_t + N_f} \quad (6)$$

The results in the Table 2, indicate that the ViT-L/14 backbone achieves the highest overall accuracy of 75.65% with a standard deviation of 3.8%, demonstrating the effectiveness of using larger vision transformer models for remote sensing scene classification. The results demonstrate a clear performance improvement with the increase in the size of the vision transformer backbone. The ViT-L/14 model outperforms the ViT-B/32 and ViT-B/16 models, suggesting that a larger model size with more parameters can capture more complex patterns and features in remote sensing images, leading to better classification accuracy. The Kappa coefficient, which measures the agreement between predicted and true classifications while accounting for chance, also shows substantial agreement across all models.

DOI: https://doi.org/10.34010/injiiscom.v6i1.14164
p-ISSN 2810-0670 e-ISSN 2775-5584

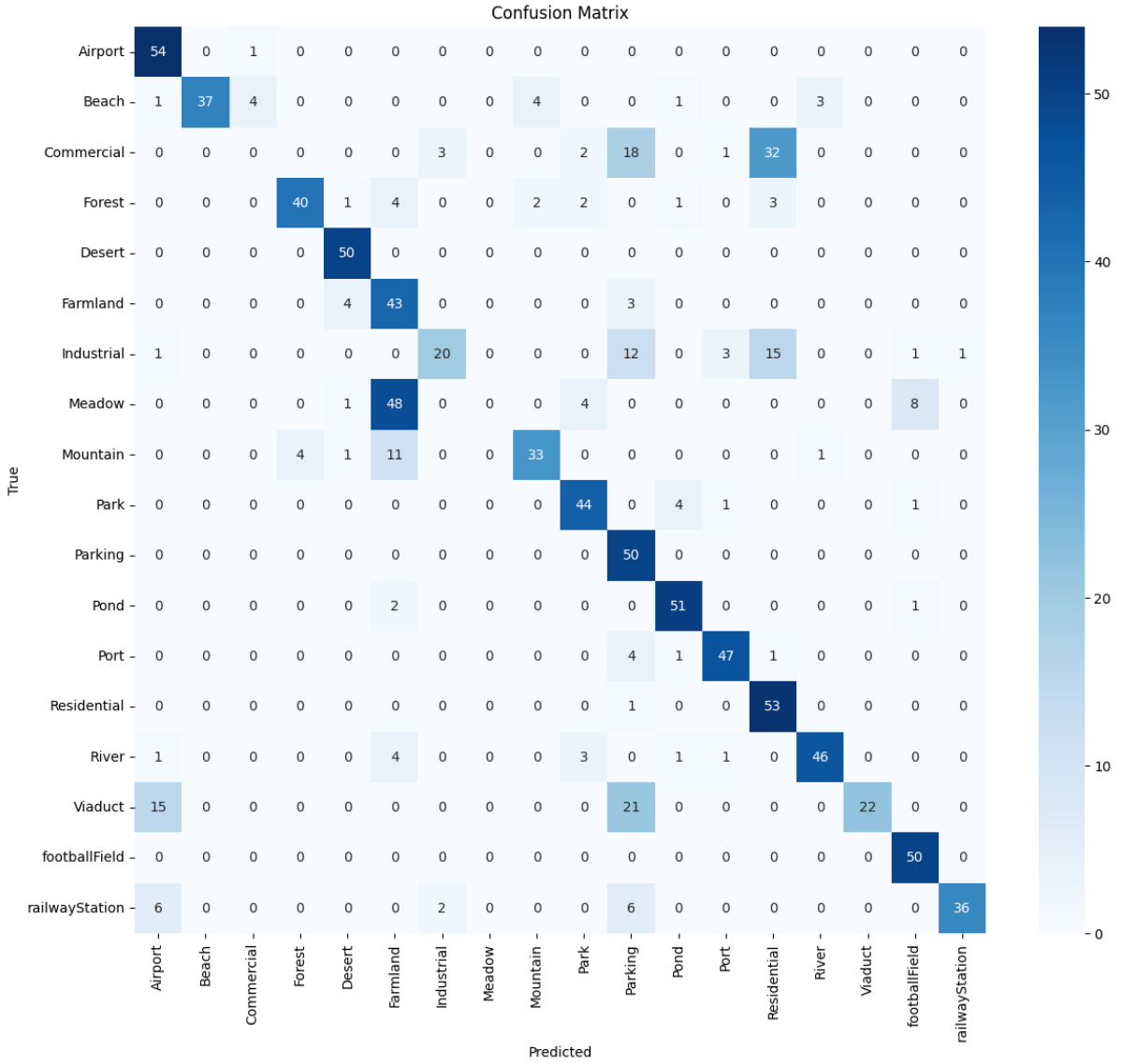

Fig. 5. Confusion Matrix of the Classification for ViT-B/32

However, it is noteworthy that the ViT-L/14 model, despite having the highest accuracy, has a slightly lower Kappa coefficient of 0.6799 compared to the ViT-B/16 model 0.7123. This discrepancy suggests that while the ViT-L/14 model performs well overall, there may be inconsistencies in its predictions for certain categories. The variation in standard deviations across the models indicates the robustness of each model. The higher SD in the ViT-B/16 model suggests more variability in its performance, while the relatively lower SD in the ViT-L/14 model indicates more consistent performance across different categories. These findings highlight the importance of model size and complexity in zero-shot remote sensing scene classification. Larger models with more parameters, like the ViT-L/14, are better suited for capturing the intricate details and variations in remote sensing images, thereby improving classification performance. Future work could focus on optimizing these models further and exploring their applicability to other datasets and classification tasks.

DOI: https://doi.org/10.34010/injiiscom.v6i1.14164
p-ISSN 2810-0670 e-ISSN 2775-5584

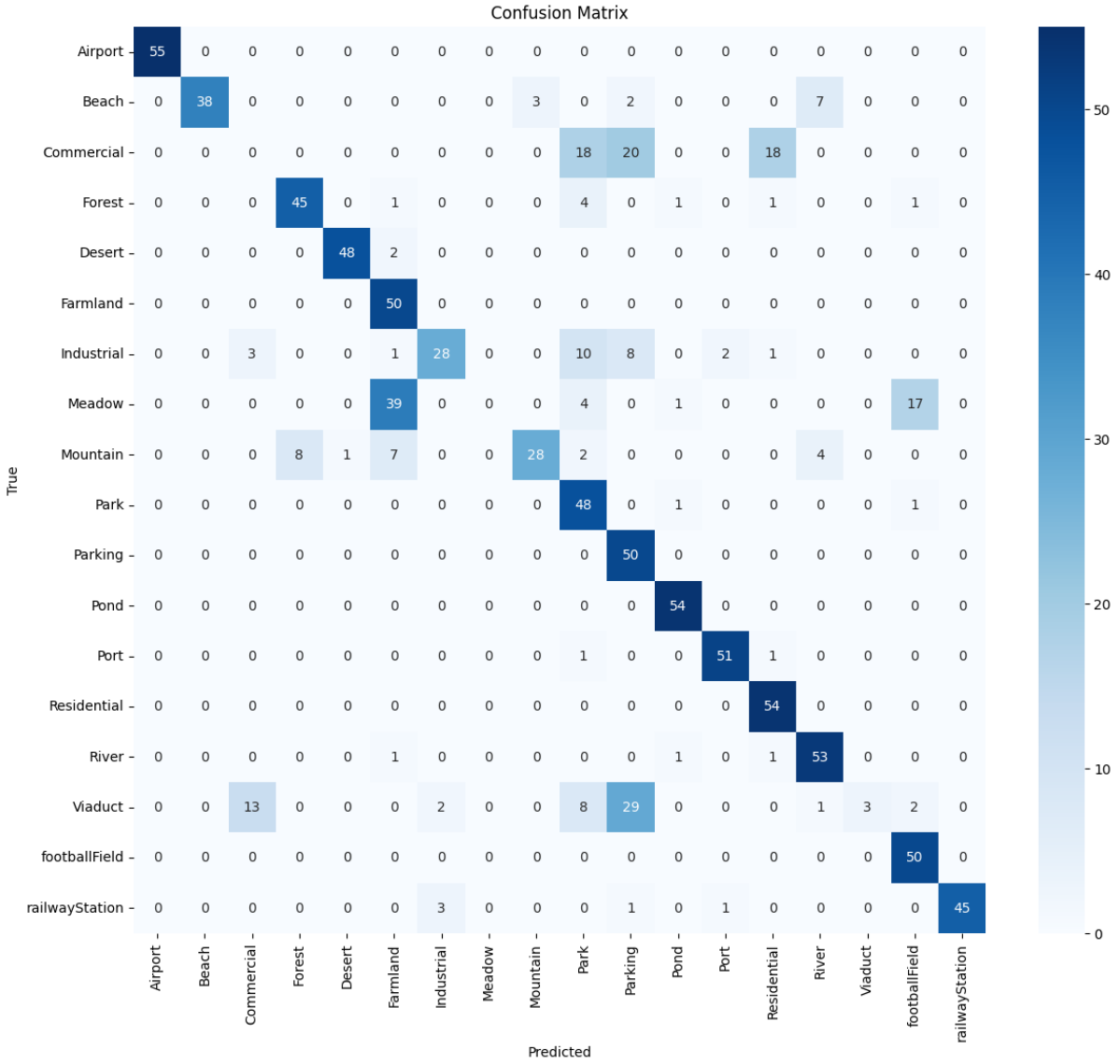

Fig. 6. Confusion Matrix of the Classification for ViT-L/14

Furthermore, we have used Confusion Matrix (CM) to analyze the performance of scene classification. As, shown in Fig. 4 provides the classification performance of the ViT-B/16 model on the WHU-RS19 dataset. The model demonstrates high accuracy in classifying certain categories such as Airport, Forest, Desert, Park, Pond, and Port, which have high values on the diagonal, 55 correctly classified for Airport, 51 for Forest, 50 for Dessert, 44 for Park, 54 for Pond, and 47 for Port. These classes likely have distinct features that the model can easily identify.

However, significant misclassifications are observed for some classes too. For instance, Commercial areas are frequently misclassified as Residential, suggesting overlapping features between these scenes. Meadow is another class with widespread misclassifications, being confused with Beach, Industrial, and Mountain, which implies variability in meadow scene features. Railway Station is misclassified with several other

DOI: https://doi.org/10.34010/injiiscom.v6i1.14164
p-ISSN 2810-0670 e-ISSN 2775-5584

classes, such as Industrial and Commercial, pointing to shared features in these scenes. Classes like Mountain present considerable challenges, with instances often being misclassified as Desert and Park, indicating difficulty in distinguishing between these natural landscapes. Industrial scenes also show confusion with Meadow and Viaduct, highlighting challenges in differentiating industrial areas from similar structures. These observations indicate that while the model performs well for classes with unique features, it faces difficulties in accurately classifying scenes with overlapping or similar visual characteristics. Then, as shown in Fig. 5, performance of the ViT-B/32, demonstrates high accuracy in classifying Residential, 50 for Parking, and 50 for FootballField have been correctly classified certain classes, such as, 54 for Airport, 50 for Dessert, 44 for Park, 51 for Pond, 47 for Port, 53 for Furthermore, as shown in Fig. 6, classification performance of the ViT-L/14, demonstrates high accuracy in recognizing classes such as, 55 for Airport, 45 for Forest, 48 for Dessert, 48 for Park, 54 for Pond, 51 for Port, 54 for Residential, 50 for Parking, and 50 for FootballField.

5. CONCLUSION

In this paper, we have demonstrated the effectiveness of a zero-shot learning (ZSL) framework for remote sensing scene classification using Vision Transformers (ViTs) and CLIP. The results show that leveraging large-scale pre-trained models and aligning visual and textual data can significantly improve the classification performance in remote sensing tasks, particularly in scenarios with imbalanced datasets. The ViT-L/14 model, in particular, exhibited superior performance, achieving the highest overall accuracy. This suggests that larger vision transformer models with more parameters are better suited for capturing the intricate details and variations in remote sensing images. Furthermore, the consistent performance across different categories underscores the robustness of our approach. In future, we hope to focus on optimizing the proposed models further to enhance their performance and efficiency. Exploring the applicability of these models to other remote sensing datasets and classification tasks will be essential for validating their generalizability. Another direction would be to investigate the potential of other advanced vision-language models and their ability to improve the alignment between visual and semantic features, thus enhancing the overall classification accuracy.

ACKNOWLEDGMENTS

We would like to thank Dr. Wang Chao, for providing us with RS dataset and helping us with the Revision.

DOI: https://doi.org/10.34010/injiiscom.v6i1.14164
p-ISSN 2810-0670 e-ISSN 2775-5584

REFERENCES

Chaib, S., H. Liu, Y. Gu, and H. Yao. 2017. 'Deep Feature Fusion for VHR Remote Sensing Scene Classification', IEEE Transactions on Geoscience and Remote Sensing, 55: 4775-84.

Chen, Feihao, and Jin Yeu Tsou. 2021. 'DRSNet: Novel architecture for small patch and low-resolution remote sensing image scene classification', International Journal of Applied Earth Observation and Geoinformation, 104: 102577.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale', ArXiv, abs/2010.11929.

Gawlikowski, J., S. Saha, A. Kruspe, and X. X. Zhu. 2022. 'An Advanced Dirichlet Prior Network for Out-of-Distribution Detection in Remote Sensing', IEEE Transactions on Geoscience and Remote Sensing, 60: 1-19.

Jia, Chao, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision." In Proceedings of the 38th International Conference on Machine Learning, edited by Meila Marina and Zhang Tong, 4904--16. Proceedings of Machine Learning Research: PMLR.

Jin, Jianhui, Wujie Zhou, Lv Ye, Jingsheng Lei, Lu Yu, Xiaohong Qian, and Ting Luo. 2022. 'DASFNet: Dense-Attention–Similarity-Fusion Network for scene classification of dual-modal remote-sensing images', International Journal of Applied Earth Observation and Geoinformation, 115: 103087.

Li, A., Z. Lu, L. Wang, T. Xiang, and J. R. Wen. 2017. 'Zero-Shot Scene Classification for High Spatial Resolution Remote Sensing Images', IEEE Transactions on Geoscience and Remote Sensing, 55: 4157-67.

Li, Y., Z. Zhu, J. G. Yu, and Y. Zhang. 2021. 'Learning Deep Cross-Modal Embedding Networks for Zero-Shot Remote Sensing Image Scene Classification', IEEE Transactions on Geoscience and Remote Sensing, 59: 10590-603.

Li, Zihao, Daobing Zhang, Yang Wang, Daoyu Lin, and Jinghua Zhang. 2022a. "Generative Adversarial Networks for Zero-Shot Remote Sensing Scene Classification." In Applied Sciences.

———. 2022b. 'Generative Adversarial Networks for Zero-Shot Remote Sensing Scene Classification', Applied Sciences, 12: 3760.

Ma, Suqiang, Chun Liu, Zheng Li, and Wei Yang. 2022. "Integrating Adversarial Generative Network with Variational Autoencoders towards Cross-Modal Alignment for Zero-Shot Remote Sensing Image Scene Classification." In Remote Sensing.

Penatti, O. A. B., K. Nogueira, and J. A. dos Santos. 2015. "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" In 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 44-51.

DOI: https://doi.org/10.34010/injiiscom.v6i1.14164
p-ISSN 2810-0670 e-ISSN 2775-5584

Quan, J., C. Wu, H. Wang, and Z. Wang. 2018. "Structural Alignment based Zero-shot Classification for Remote Sensing Scenes." In 2018 IEEE International Conference on Electronics and Communication Engineering (ICECE), 17-21.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. "Learning Transferable Visual Models From Natural Language Supervision." In Proceedings of the 38th International Conference on Machine Learning, edited by Meila Marina and Zhang Tong, 8748--63. Proceedings of Machine Learning Research: PMLR.

Romera-Paredes, Bernardino, and Philip H. S. Torr. 2017. 'An Embarrassingly Simple Approach to Zero-Shot Learning.' in Rogerio Schmidt Feris, Christoph Lampert and Devi Parikh (eds.), Visual Attributes (Springer International Publishing: Cham).

Wang, C., J. Li, A. Tanvir, J. Yang, T. Xie, L. Ji, and T. Zhang. 2024. 'Zero-Shot Remote Sensing Scene Classification Method Based on Local-Global Feature Fusion and Weight Mapping Loss', IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17: 2763-76.

Wang, C., G. Peng, and B. De Baets. 2021. 'A Distance-Constrained Semantic Autoencoder for Zero-Shot Remote Sensing Scene Classification', IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14: 12545-56.

Xian, Y., Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. 2016. "Latent Embeddings for Zero-Shot Classification." In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 69-77.

Zhu, Junjie, Ke Yang, Naiyang Guan, Xiaodong Yi, and Chunping Qiu. 2023. 'HCPNet: Learning discriminative prototypes for few-shot remote sensing image scene classification', International Journal of Applied Earth Observation and Geoinformation, 123: 103447.

Zhu, Qiqi, Yang Lei, Xiongli Sun, Qingfeng Guan, Yanfei Zhong, Liangpei Zhang, and Deren Li. 2022. 'Knowledge-guided land pattern depiction for urban land use mapping: A case study of Chinese cities', Remote Sensing of Environment, 272: 112916.

DOI: https://doi.org/10.34010/injiiscom.v6i1.14164
p-ISSN 2810-0670 e-ISSN 2775-5584