Enhancing Remote Sensing Scene Classification with Channel-Spatial CNN (CS-CNN)

S. Akila Agnes[1], Bhargavi Pedada, Raju Sambangi, Mohitha Dasari, Vijay Prakash Penugonda & Sai Ram Pati

Department of Computer Science and Engineering, GMR Institute of Technology, Andhra Pradesh, India.

Abstract Remote sensing scene classification plays a vital role in analysing Earth's surface using satellite or aerial imagery. The classification of remote images unlocks a world of applications, including Land Use Planning, Urban Development, Environmental Monitoring, Disaster Management, Defence, and Geospatial Analysis. The standard Convolutional Neural Networks (CNNs) often struggle to extract class-relevant features from high-resolution remote sensing images. In response to this challenge, our research introduces an innovative approach: the integration of a CNN with a channel and spatial attention mechanism, aiming to achieve precise classification of remote scene images. Unlike conventional CNNs, which primarily extract fundamental features, our proposed approach aims to extract both channel-based and spatial-based features from the respective images to achieve improved performance. The proposed Channel-Spatial Convolutional Neural Network (CS-CNN) model automatically categorizes the RSSCN7 dataset images distinct classes, including Grass, Field, Industry, River Lake, Forest, Residential, and Parking lot. This model enhances the accuracy while concurrently diminishing loss in the classification system. It accomplishes this feat by directing its attention to pertinent regions within the images and effectively capturing long-range dependencies inherent in remote sensing imagery. To assess the efficacy of our proposed method, we conducted an extensive evaluation involving multiple deep learning models, encompassing standard CNNs, CNNs enriched with channel and spatial attention mechanisms, ResNet50, ResNet50 with channel attention, VGG19, and VGG19 with attention mechanisms. Experimental results demonstrate the superiority of the CNN integrated with both channel and spatial attention mechanisms, achieving an impressive accuracy rate of 96% in the classification of remote images.

[1]Dr. S. Akila Agnes, Assistant Professor, Department of Computer Science and Engineering, GMR Institute of Technology, Andhra Pradesh, India.-mail: akila.s@gmrit.edu.in

Keywords : CNN, Remote scene classification, Channel attention, RSSCN7, Spatial attention

1. Introduction

Remote sensing scene classification is a crucial task in the field of Earth observation and satellite imagery analysis. It involves the automatic categorization of images captured by remote sensing platforms, such as satellites or aerial vehicles, into predefined classes or land cover types [1]. This classification process plays a vital role in various applications, including environmental monitoring, urban planning, agriculture, disaster assessment, and more. Since the convolutional neural network (CNN) can learn features, it has improved accuracy in the domain of classifying remote images. However, these CNN-based models for classifying images have a lot of deep layers that don't quite describe how the objects in the image relate to one another [2]. However, the majority of current methods for remote sensing scene classification only focus on global information; categories of remote images depend on regions containing class-specific ground objects. This causes redundant data and poor remote image scene classification. To overcome the drawbacks of current methods, a convolutional neural network with an attention mechanism is needed [3]. CNN model integrating with attention mechanism have revolutionized image classification tasks due to their ability to automatically learn both channel-based and spatial-based features from the input data. This model is particularly well-suited for remote sensing scene classification as it can effectively handle the large-scale and complex spatial patterns present in satellite imagery.

Attention Mechanism is an extension to the traditional CNN architecture that aims to enhance the model's capability to focus on the most relevant regions of the input image while suppressing irrelevant or noisy information. Attention mechanisms help to improve the discriminative power of CNNs by allowing them to dynamically adapt their focus on different parts of the scene, leading to more accurate and robust classification results [4]. The integration of attention mechanisms with CNNs in remote sensing scene classification tasks has shown promising results. By leveraging attention, the model can selectively emphasize informative regions in the image, such as roads, buildings, vegetation, water bodies, etc., while downplaying less relevant areas, like clouds or shadows. This selective attention can significantly improve the model's performance, especially in challenging scenarios with complex backgrounds or when dealing with partial occlusions [5].

Transfer learning techniques such as Alex Net and VGG are used for the remote sensing scene classification and achieved an accuracy of 93% and 92% respectively. Compared with the traditional methods like convolutional neural networks these pre-trained models got a good performance in this image classification task. However, these models failed to extract the region-specific features from the input images which affected the overall performance of the model by leading the model to misclassification of images [6-7]. Applying attention mechanisms to transfer

learning models offers several advantages and disadvantages. On the positive side, attention mechanisms enhance the model's ability to focus on relevant information in the source data while transferring knowledge to a new task or domain. This results in improved model performance and generalization, as it can selectively attend to important features and relationships. Additionally, attention-based transfer learning models can adapt to varying input lengths and capture long-range dependencies, making them versatile for various tasks. However, there are some drawbacks to consider. Attention mechanisms increase model complexity and computational requirements, potentially leading to longer training times and higher resource demands. Moreover, they might suffer from overfitting if not carefully regularized. Finally, selecting the appropriate attention architecture and tuning hyperparameters can be challenging, requiring expertise and careful experimentation. In summary, attention-based transfer learning models offer substantial benefits in terms of performance and adaptability but come with trade-offs related to complexity and resource utilization [8-10].

In this work, the proposed CNN with channel and spatial attention mechanism model overcomes all existing challenges till now in remote sensing scene classification. CNN is well known for image classification but it can't extract more class-specific features from high-resolution images in such domains as remote images. So, an attention mechanism is required to be added in between the convolutional neural network to capture the channel-wise and spatial-wise features of the input image [11-13].

The rest of the paper follows this structure. Section 2 describes the CNN with attention mechanism architecture and remote sensing scene image classification method. Section 3 presents the outcomes and experiments conducted using the proposed approach on the RSSCN7 dataset. Section 4 concludes the proposed work.

2. Methodology

Established models such as CNNs may encounter limitations in extracting an increased number of class-specific features from high-resolution images. To overcome this challenge, the incorporation of attention layers within the convolutional neural network can offer a potential remedy. In this study, various deep learning models were trained on the RSSCN7 dataset, including CNN, CS-CNN (Channel-Spatial Convolutional Neural Network, which integrates both channel and spatial attention mechanisms), ResNet50, ResNet50 with channel attention (ResNet50-CA), VGG19, and VGG19 with attention mechanism (VGG19-CA). Fig.1 illustrates the integration of attention mechanisms into the CNN architecture.

Channel-Spatial Convolutional Neural Network (CS-CNN)

The proposed CS-CNN model is a deep learning architecture developed using convolutional neural networks and an attention mechanism. It comprises four convolutional layers with a 3x3 kernel size, followed by four attention and max-pooling layers. Then there is a classification block that contains two dense layers followed by a dropout of 25%. Then Softmax activation function is applied to classify the input images into one of the predefined class labels.

The Convolutional Neural Network (CNN), a deep learning model originally tailored for image and spatial data analysis, has undergone a significant enhancement. While it traditionally relies on specialized convolutional layers to automatically extract pertinent features from input data [14-15], in this context, the resized image (224x224x3) sourced from the RSSCN7 dataset is channeled through a convolutional layer featuring 32 filters, each with a 3x3 filter size.

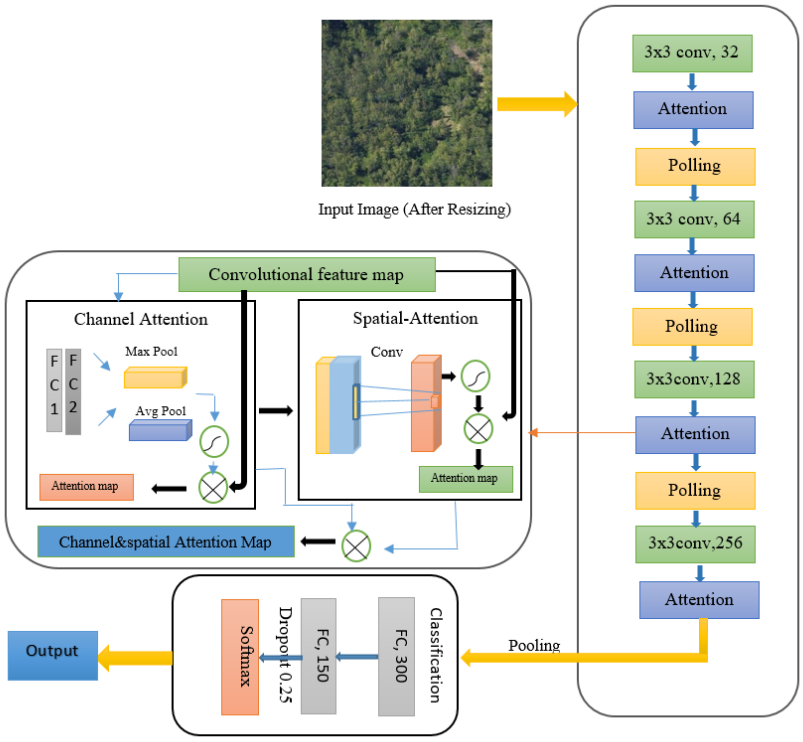

Fig 1. Architecture of channel-spatial convolutional neural network (CS-CNN) for remote sensing scene classification.

This convolutional layer generates a feature map, which notably serves as the input for the subsequent Attention layer, representing a crucial modification to the conventional CNN architecture. Adding attention layers in between the convolutional layers in this model helps to increase the performance of remote sensing scene image classification. The attention mechanism integrated into this model in-

cludes two blocks such as channel attention and spatial attention. This block takes the convolutional feature map (F) as the input and produces the attention map (F') as the output. These attention blocks include two sub blocks such as channel attention and spatial attention.

Channel attention mechanism, a crucial component in CNN for feature enhancement. It begins by detecting the appropriate channel axis based on the data format. Two shared dense layers are created as shown in the architecture (Fig.1) to calculate attention weights for each channel, aiming to capture channel-wise dependencies. The code then applies global average and max pooling to the input feature maps to aggregate information across spatial dimensions. These pooled representations are reshaped and processed through the shared layers to compute attention scores. The results from average and max pooling are combined through addition and passed through a sigmoid activation to obtain the attention weights, which determine the importance of each channel. If necessary, the data format is adjusted, and finally, the input feature maps are scaled by these attention weights, enhancing the representation of relevant channels while suppressing less informative ones. This channel attention mechanism allows the CNN to adaptively focus on the most critical information within the feature maps and this output feature map ($M_c(F)$) is given as the input to the next sub-block i.e. spatial attention block.

$$M_c(F) = \sigma\left(Avg\_pool\,(F) + Max\_pool(F)\right) \otimes F \qquad \text{Eq. 1}$$

The channel attention map ($M_c(F)$) is computed using equation 1and the input convolutional feature map (F) is provided as the input to the spatial attention block for enhancing specific spatial regions in an input feature map. It begins by determining the input feature's channel format. The channel format check ensures compatibility between data formats. Then it computes the average and maximum values along the channel dimension, creating two attention maps. These maps are concatenated to form a two-channel attention feature. A convolutional layer with a sigmoid activation is applied to the concatenated feature map, producing a spatial attention map that emphasizes relevant regions. Finally, the input feature is element-wise multiplied with the attention map, effectively amplifying the importance of certain spatial locations based on their relevance, to produce a spatial attention map ($M_s\,(F)$) using equation 2.This spatial attention block aims to capture both global and local context information in the input feature map.

$$M_s\,(F) = \sigma\left(f^{7\times7}\left(Avg\_pool(F); Max\_pool(F)\right)\right) \otimes F \qquad \text{Eq. 2}$$

The outputs of the above two sub-attention blocks[$(M_c(F), M_s(F)$] are then multiplied to get the overall channel and spatial attention feature map ($F^{`}$) using equa-

tion 3, which is then fed to the next pooling layers. The process continues for further layers and finally classifies the given input into one of the seven categories.

$$F' = M_c(F) \otimes M_s(F') \quad \text{Eq. 3}$$

The proposed model is evaluated against the test data using various performance metrics, including Accuracy, Precision, Recall, and F1 Score. The outcomes of the proposed model are detailed in subsequent sections.

RSSCN7 Dataset:

The RSSCN7 dataset, made publicly available by Wuhan University in 2015, is used to train and evaluate the suggested deep learning models. This dataset consists of satellite images obtained from Google Earth that were originally collected for remote sensing scene classification tasks. The seven typical scene categories are grass, field, industry, river-lake, forest, and parking lot. The entire dataset consists of 2800 images, with 400 high-resolution RGB images measuring 256 by 256 pixels in each category. Sample images from each class in the RSSCN7 dataset are shown in figure 2. Significant amounts of training data are needed for deep learning models to function well. So, the available RSSCN7 dataset is augmented to produce more image samples by using the ImagedataGenerator module from the Keras framework. Each image is then resized to 224X224 pixels to reduce the computational load on the deep learning models.

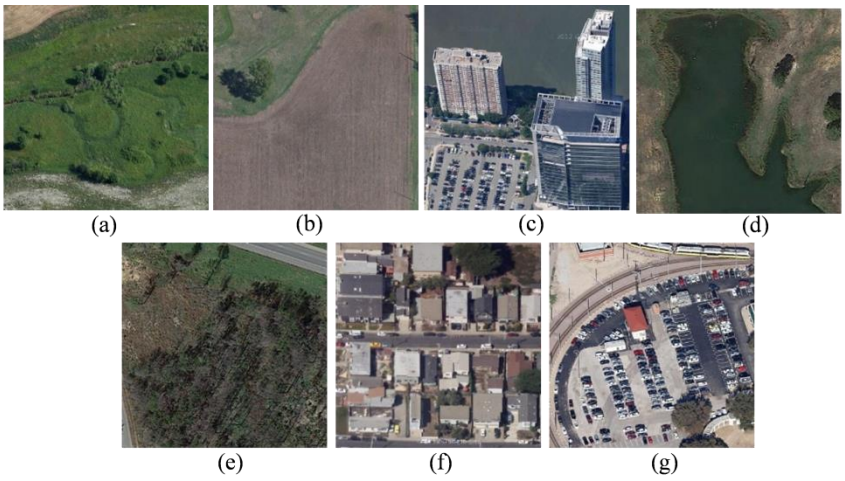

Fig 2. Sample Images form RSSCN7 dataset: (a) Grass, (b) Field, (c) Industry, (d) River Lake, (e) Forest, and (g) Parking Lot

3. Results and Discussions

Deep learning models such as CNN, CS-CNN, ResNet50, ResNet50-CA, VGG19, and VGG19-CA are applied on the benchmark RSSCN7 dataset to classify the remote sensing scene images. Among the above-mentioned models, our proposed model CS-CNN i.e. CNN with channel and spatial attention mechanism outperformed the baseline models and achieved an accuracy of 96% with less loss of 0.1. Table 1 discusses a detailed comparison of different deep learning models based on parameters such as Accuracy, Training loss, No. of trainable parameters and Time per step in milliseconds.

Table 1. Performance Comparison of Various Deep Learning Models for Remote Sensing Scene Classification

Model	Accuracy	Training Loss	Trainable Parameters	Time per step (ms)
CNN	84%	0.434	35,163,847	734
VGG19	94%	0.131	20,024,384	221
ResNet50	90%	0.250	26,043,601	600
VGG19-CA	88%	0.620	22,563,879	553
ResNet50-CA	91%	0.230	25,648,647	777
CS-CNN(proposed)	96%	0.100	15,489,943	46

Performance Analysis of Various Models in RS-Scene Classification:

To determine which deep learning architectures are most suitable for this specific task, it is essential to analyze the performance of various models in remote sensing scene classification. In this section, we present different classification reports generated by these deep learning models. A classification report holds significant importance in deep learning as it offers a comprehensive evaluation of a model's performance across distinct classes. It encompasses key metrics such as precision, recall, support, and F1-Score, which assist in assessing a model's accuracy and its capability to distinguish between different classes. This report plays a vital role in comprehending where a model excels and where it may need improvements, thereby guiding model refinement and informed decision-making.

Table 2 presents the classification report for ResNet50 and ResNet50-CA on the RSSCN7 dataset for remote sensing scene classification. ResNet50-CA consistently enhances recall for the majority of classes, suggesting that the Channel Attention mechanism improves class recognition. The majority of F1-Scores for ResNet50-CA are equal or greater, demonstrating its ability to balance precision and recall. ResNet50-CA's weighted average F1-Score is 0.85, surpassing ResNet50's 0.81, indicating superior overall performance. Compared to ResNet50, ResNet50-CA performs significantly better, with an accuracy of 0.85. However, ResNet50's

performance is not consistent across classes; some classes have lower precision and recall than others.

Table 2. Classification Report of ResNet50 and ResNet50-CA on RSSCN7 Dataset for Remote Sensing Scene Classification

Class label	ResNet50			ResNet50-CA		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Grass	0.82	0.71	0.76	0.82	0.88	0.85
Field	0.67	0.98	0.79	0.93	0.79	0.85
Industry	0.69	0.83	0.76	0.69	0.87	0.77
River Lake	0.93	0.82	0.87	0.89	0.93	0.91
Forest	0.92	0.78	0.85	0.93	0.90	0.91
Resident	0.85	0.80	0.82	0.90	0.97	0.93
Parking	0.91	0.74	0.81	0.88	0.66	0.75
Accuracy			0.81			0.85
Macro avg	0.83	0.81	0.81	0.86	0.85	0.85
Weighted avg	0.83	0.81	0.81	0.86	0.85	0.85

Table 3 presents the classification report for two different models, VGG19 and VGG19-CA, on the RSSCN7 dataset for remote sensing scene classification. The addition of the Channel Attention mechanism (CA) to VGG19 has led to mixed results across different classes. In some cases, such as the "Field" and "Parking" classes, CA has improved recall but reduced precision, resulting in lower F1-Scores.In other cases, like the "Grass" and "Resident" classes, CA has negatively impacted both precision and recall, leading to lower F1-Scores.The weighted average F1-Score for VGG19-CA is 0.80, which is slightly lower than the F1-Score of 0.89 for VGG19. This suggests that, in this particular scenario, the inclusion of the Channel Attention mechanism did not significantly improve the overall performance. The accuracy for VGG19-CA is also lower at 0.80 compared to 0.89 for VGG19.

Table 3. Classification Report of VGG19 and VGG19-CA on RSSCN7 Dataset for Remote Sensing Scene Classification

Class label	VGG19			VGG19-CA		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Grass	0.88	0.86	0.87	0.72	0.75	0.74
Field	0.90	0.92	0.91	0.75	0.90	0.82
Industry	0.80	0.81	0.81	0.79	0.70	0.74
River Lake	0.89	0.96	0.92	0.90	0.80	0.85
Forest	0.98	0.88	0.93	0.84	0.80	0.82
Resident	0.95	0.97	0.96	0.92	0.74	0.82
Parking	0.86	0.85	0.86	0.74	0.91	0.82
Accuracy			0.89			0.80
Macro avg	0.89	0.89	0.89	0.81	0.80	0.80
Weighted avg	0.89	0.89	0.89	0.81	0.80	0.80

Table 4 presents the classification report for two different models, CNN and CS-CNN (Proposed), on the RSSCN7 dataset for remote sensing scene classification. CS-CNN (Proposed) consistently outperforms CNN across all classes, significantly improving precision, recall, and F1-Scores. The weighted average F1-Score for

CS-CNN is 0.90, which is substantially higher than the F1-Score of 0.53 for CNN. This indicates that CS-CNN outperforms CNN in terms of overall F1-Score. The accuracy for CS-CNN is also significantly higher at 0.90 compared to 0.56 for CNN, emphasizing the substantial positive impact of CS-CNN. The results demonstrate the model's strong performance in classifying remote sensing scenes, exhibiting high precision, recall, and F1-Score values across most classes. However, it's worth noting that the "Parking" class exhibits slightly lower performance metrics. Specifically, the "Parking" class has a precision of 0.91, a recall of 0.86, and an F1-Score of 0.89. These metrics are marginally lower than those of other classes, suggesting that some instances of the "Parking" class may have been misclassified as something else, leading to the reduced recall for this class. The weighted average metrics suggest that this performance is consistent across the entire dataset, with an overall accuracy of 96%. CS-CNN (Proposed) consistently outperforms all these models in terms of precision, recall, F1-Score, and accuracy across various classes. It achieves a substantially higher overall accuracy and F1-Score, making it the top-performing model in this comparison.

Table 4. Classification report of CNN and CS-CNN on RSSCN7 Dataset for Remote Sensing Scene Classification

Class label	CNN			CS-CNN(Proposed)		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Grass	0.40	0.83	0.56	0.93	0.90	0.92
Field	0.40	0.85	0.55	0.96	0.96	0.96
Industry	0.61	0.80	0.69	0.96	0.95	0.96
River Lake	0.55	0.54	0.54	0.97	0.96	0.97
Forest	0.71	0.70	0.70	0.96	0.91	0.94
Resident	0.88	0.63	0.63	0.95	0.94	0.95
Parking	0.54	0.55	0.55	0.91	0.86	0.89
Accuracy			0.56			0.96
Macro avg	0.59	0.57	0.53	0.96	0.96	0.96
Weighted avg	0.58	0.56	0.53	0.96	0.96	0.96

Model Comparison with Confusion Matrices:

Below are the different confusion matrix representations of the deep learning models. In machine learning, a confusion matrix is essential because it provides a thorough analysis of a model's performance by displaying false positives, false negatives, true positives, and true negatives. This matrix helps assess the accuracy and robustness of a model, especially when class imbalances or misclassification costs are at play. It provides valuable insights for making informed decisions regarding model adjustments and fine-tuning. The confusion matrices for various models, including CNN, ResNet50, ResNet50-CA, VGG19-CA, VGG19, and CS-CNN (Proposed), are shown in Fig. 3. In this confusion matrix, you have the actual class labels on the left (rows) and the predicted class labels on the right (columns). The numbers in each cell represent the counts of instances. The confusion matrix provides insights into how well the model performed in classifying different classes. The CS-CNN (Proposed) model seems to have classified more true positives compared to the actual classes, indicating its effectiveness in identifying

those categories. However, there are classes like "Parking" where the model has a higher number of false negatives, indicating room for improvement in classification accuracy for that class.

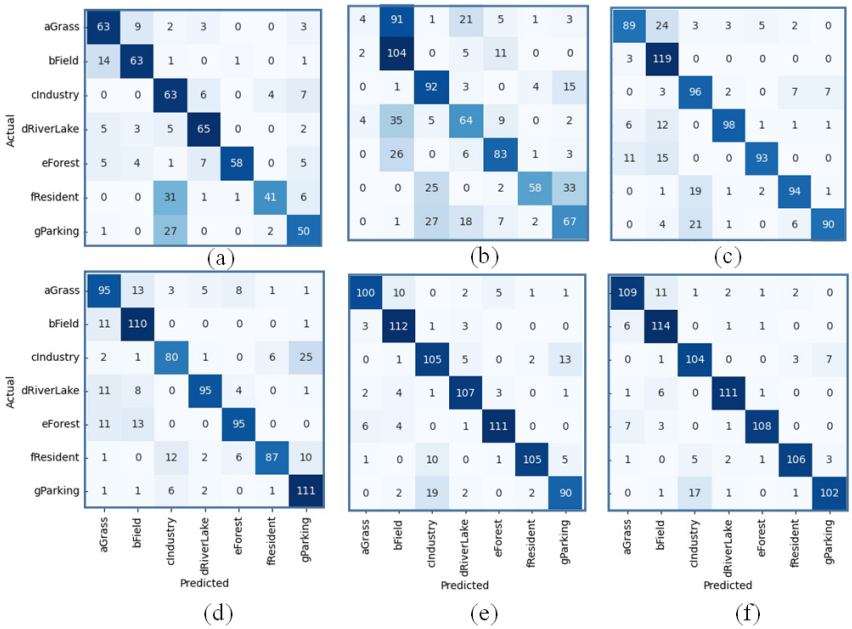

Fig 3. Confusion matrix of (a) CNN, (b) ResNet50, (c) ResNet50-CA, (d) VGG19-CA, (e) VGG19, and (f) CS-CNN (Proposed), displaying the actual vs. predicted classes.

4. Conclusion

The proposed work resolves the challenges of remote sensing scene classification such as remote sensing data can exhibit complex spatial and spectral variations so deep learning models need to effectively capture these intricate patterns, which can be challenging with traditional methods. However, our proposed model CS-CNN overcomes this by focusing on the more relevant information from images by using an attention mechanism both channel-wise and spatial-wise and achieves an accuracy of 96%. Deep learning models like ResNet50, CNN, VGG19, CNN with Attention, VGG19 with Attention, and ResNet50 with Attention have shown promising results in remote sensing scene classification. Among these models, the CS-CNN model effectively addresses complex challenges in remote sensing scene classification and has led to significant improvements in accuracy, precision, recall, and F1-score compared to other models, especially those without attention mechanisms. With a significant emphasis on accuracy, efficiency, and ethical considerations, he future of research in remote sensing scene classi-

fication will unite cutting-edge attention mechanisms, including Transformer Attention and Multi-Head Attention and maintain a dedicated focus on practical, real-world applications.

References

[1] Zhao, M., Meng, Q., Zhang, L., Hu, X., & Bruzzone, L. (2023). Local and long-range collaborative learning for remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*.

[2] Tong, W., Chen, W., Han, W., Li, X., & Wang, L. (2020). Channel-attention-based DenseNet network for remote sensing image scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 4121-4132.

[3] Li, F., Feng, R., Han, W., & Wang, L. (2020). An augmentation attention mechanism for high-spatial-resolution remote sensing image scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 3862-3878.

[4] Chen, S. B., Wei, Q. S., Wang, W. Z., Tang, J., Luo, B., & Wang, Z. Y. (2021). Remote sensing scene classification via multi-branch local attention network. *IEEE Transactions on Image Processing*, *31*, 99-109.

[5] Thirumaladevi, S., Swamy, K. V., & Sailaja, M. (2023). Remote sensing image scene classification by transfer learning to augment the accuracy. *Measurement: Sensors*, *25*, 100645.

[6] Cao, R., Fang, L., Lu, T., & He, N. (2020). Self-attention-based deep feature fusion for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, *18*(1), 43-47.

[7] Song, S., Yu, H., Miao, Z., Zhang, Q., Lin, Y., & Wang, S. (2019). Domain adaptation for convolutional neural networks-based remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, *16*(8), 1324-1328.

[8] Zhang, J., Zhao, H., & Li, J. (2021). TRS: Transformers for remote sensing scene classification. *Remote Sensing*, *13*(20), 4143.

[9] Shabbir, A., Ali, N., Ahmed, J., Zafar, B., Rasheed, A., Sajid, M., ... & Dar, S. H. (2021). Satellite and scene image classification based on transfer learning and fine tuning of ResNet50. *Mathematical Problems in Engineering*, *2021*, 1-18.

[10] Wang, X., Duan, L., Ning, C., & Zhou, H. (2021). Relation-attention networks for remote sensing scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 422-439.

[11] Cheng, G., Xie, X., Han, J., Guo, L., & Xia, G. S. (2020). Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 3735-3756.

[12] Adegun, A. A., Viriri, S., & Tapamo, J. R. (2023). Review of deep learning methods for remote sensing satellite images classification: experimental survey and comparative analysis. *Journal of Big Data*, *10*(1), 93.

[13] Xu, X., Chen, Y., Zhang, J., Chen, Y., Anandhan, P., & Manickam, A. (2021). A novel approach for scene classification from remote sensing images using deep learning methods. *European Journal of Remote Sensing*, *54*(sup2), 383-395.

[14] Kong, J., Gao, Y., Zhang, Y., Lei, H., Wang, Y., & Zhang, H. (2021). Improved attention mechanism and residual network for remote sensing image scene classification. *IEEE Access*, *9*, 134800-134808.

[15] Wenmei Li, Ziteng Wang, Yu Wang, Jiaqi Wu, Juan Wang, Yan Jia, and Guan Gui (2020) Classification of High-Spatial-Resolution Remote Sensing Scenes Method Using Transfer Learning and Deep Convolutional Neural Network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*,1986 – 1995.