The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W5-2025
13th International Conference on Mobile Mapping Technology (MMT 2025)
"Mobile Mapping for Autonomous Systems and Spatial Intelligence", 20–22 June 2025, Xiamen, China

MultiTrans-LC: Multimodal Fusion Transformer for Remote Sensing Land Cover Classification

Qixuan Wang[1], Ning Li[*1,2], Yiheng Chen[1], Hainiu Zhu[1]

[1] College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, 211106 Nanjing, China - (wangqx, lnee, cyh117 , zhn123)@nuaa.edu.cn;
[2] Key Laboratory of Radar Imaging and Microwave Photonics (Nanjing University of Aeronautics and Astronautics), Ministry of Education, 211106 Nanjing, China;

Keywords: Multimodal Fusion, Transformer, Remote Sensing, Land Cover Classification.

Abstract

The use of remote sensing images for land cover classification is crucial for environmental monitoring, urban planning, and sustainable resource management. Despite advances in deep learning, existing methods suffer from blurred boundaries in complex landscapes and perform poorly in identifying small or overlapping land cover categories. This article introduces MultiTrans LC, a novel multimodal fusion framework that integrates visual language interaction and boundary perception optimization to address these challenges. The proposed architecture utilizes a hierarchical Transformer encoder to extract global visual features from high-resolution images and aligns them with semantic embeddings in text prompts through cross modal attention. The visual language decoder further refines the multi-scale feature representation through progressive fusion, while the edge aware loss function jointly optimizes pixel level classification and boundary localization. Experiments on three benchmark datasets (GID-15, LoveDA, RSSCN7) have demonstrated state-of-the-art performance, achieving an overall accuracy of **90.7%** and a Kappa coefficient of **0.901** on GID-15, which is **1.6%** higher than the leading method in OA. Visualization confirms that MultiTrans LC performs well compared to CNN and Transformer baselines. By bridging visual and textual semantics, MultiTrans LC improves the accuracy of large-scale land cover mapping and provides a powerful solution for geospatial intelligence applications. Discussed the limitations and future directions of open vocabulary classification and edge device deployment.

1. Introduction

Land cover classification is a technical process that identifies the surface cover type of each pixel in remote sensing imagery to generate comprehensive thematic maps. As a critical component of foundational geospatial data, land cover maps provide essential spatiotemporal change information for major applications such as urban planning, dynamic monitoring of natural disasters, and ecological vulnerability assessments. With advancements in satellite sensor technology and the maturation of unmanned aerial vehicle photogrammetry, high-resolution remote sensing imagery, characterized by rich spatial and textural features, has emerged as a primary data source for land cover classification studies.

In recent years, breakthroughs in deep learning have significantly advanced the practical implementation of computer vision tasks, including image classification, object detection, and semantic segmentation. Unlike traditional methods that rely on manually engineered features with limited generalizability, deep learning approaches autonomously extract discriminative hierarchical feature representations through large-scale annotated datasets. Current research predominantly focuses on optimizing convolutional neural network (CNN) architectures (O'shea and Nash, 2015) and exploring Transformer-based models (Kalyan et al., 2021). However, existing methodologies primarily concentrate on single-modal image data analysis, failing to fully exploit the complementary semantic information embedded in multimodal data (e.g., optical imagery paired with LiDAR point clouds, multispectral bands, or textual descriptions). This limitation constitutes a key bottleneck in improving land cover classification accuracy. Consequently, developing a foundational framework for synergistic multimodal data fusion has become a vital research direction to overcome current technical constraints.

In this paper, we introduce the Transformer model into the semantic segmentation task of remote sensing images and propose a multi-modal change detection framework, namely MultiTrans-LC. In the context of multi-modal change detection tasks, changes are typically identified using a single modality, primarily remote sensing images. To integrate text-based cues, we utilize the Transformer model to generate descriptive prompts for common land cover category classification. During the image encoding stage, we construct a Transformer network to extract image features from high-resolution images. During the text encoding stage, we apply a Transformer network to extract text features from text prompts. The main contributions of this paper are as follows:

(1) We designed a multi-modal decoder based on Transformer to enhance the semantic relationships between image-text feature pairs.

(2) The proposed MultiTrans-LC achieves state-of-the-art performance on the GID15, LoveDA and RSSCN7 datasets.

The rest of this paper is organized as follows: Section II describes the related algorithms; In Section III, we introduce the overall framework of MultiTrans-LC; Section IV compares MultiTrans-LC with other advanced algorithms through experiments, confirming its effectiveness; and Section V summarizes the paper.

2. Related Works

2.1 Semantic segmentation of remote sensing images

Compared to traditional manual ground surveys, remote sensing technology has become the mainstream means of sur-

face monitoring due to its advantages such as wide coverage, fast acquisition speed, and rich information. It is extensively used in soil research , geological engineering , land resources , and other fields. The quality of remote sensing images has increased along with the rapid development of the technology. Remote sensing images can provide a wealth of information about ground objects, such as ground vegetation cover, ground temperature, and land use.Semantic segmentation of remote sensing images is a key step in understanding their content. By converting pixel level information into interpretable land cover categories, it provides structured support for subsequent applications.As a result, the semantic segmentation technique for remote sensing images has significant research implications.

Most traditional machine learning-based remote sensing image interpretation algorithms adopt feature extraction and feature analysis, and the interpretation effect is good for specific scenes and datasets. However, classic machine learning algorithms have restricted feature extraction and cannot accurately capture the nuances of the input. When the background level of the remote sensing image to be processed is complicated and the target scale has large fluctuations, the model accuracy suffers and under-fitting or over-fitting occurs.(Wang et al., 2024)

2.2 Multimodal representation learning

Multimodal Representation Learning is an important research direction in the field of machine learning. It aims to extract and integrate effective information from data of various modalities, such as text, images, audio, and video, to generate unified and interpretable representations that support downstream tasks like classification, retrieval, and generation. The core objective is to model the correlations and complementarities between different modalities, thereby enhancing the model's ability to understand complex scenarios.

2.3 Transformer backbones

ViT (Yuan et al., 2021) is the first work to prove that a pure Transformer can achieve state-of-the-art performance in image classification. ViT treats each image as a sequence of tokens and then feeds them to multiple Transformer layers to make the classification.

DETR (Zheng et al., 2023) is the first to use Transformers for end-to-end object detection framework without non-maximum suppression (NMS). Other works have also used Transformers for semantic segmentation in tasks such as tracking, super-resolution, re identification, coloring, retrieval, and multimodal learning.

3. Method

3.1 Overall Architecture

In this study,we introduce MultiTrans-LC is a multimodal Transformer network designed for remote sensing land cover classification.As illustrated in Figure 1, the framework comprises three core components:Multimodal Feature Extraction,Cross-modal Fusion Module and Vision-Language-Driven Decoder.In the first component,it captures visual features from remote sensing imagery and semantic embeddings from language-text descriptions.For the second component, it aligns and fuses heterogeneous features through attention mechanisms.In the last part, we make full use of the vision-language features from the encoding stage.We introduce graphic text joint features in the decoding stage, and fuse the visual language representation obtained in the encoding stage with the decoding features to construct a semantic enhanced decoder, in order to improve the recognition accuracy of the model for complex terrain or changing areas.

3.2 Multimodal Feature Extraction

Nowadays, the integration of computer vision and natural language processing is becoming increasingly tight, giving rise to numerous outstanding projects that combine these two fields.

3.2.1 Visual Feature Encoder: In this study, we adopt a hierarchical Transformer-based encoder to extract visual representations from remote sensing imagery. We adopt a Transformer based encoding structure to divide remote sensing images into image blocks and model their global contextual relationships. At the same time, we perform semantic encoding on text descriptions to achieve alignment and complementarity between images and language in the feature space. This design enables the network to capture global contextual cues, which is particularly beneficial for land cover classification in complex and heterogeneous landscapes.

3.2.2 Textual Feature Encoder: For the linguistic modality, we utilize a Transformer-based text encoder to process semantic descriptions of land cover categories. Each textual prompt (e.g., "dense urban area" or "deciduous forest") is tokenized and mapped into embedding vectors. These embeddings are refined through multi-head self-attention mechanisms, allowing the model to capture intra-sentence relationships and generate semantically rich representations. The resulting textual features serve as high-level semantic anchors for subsequent cross-modal alignment.

3.3 Multimodal Fusion Module

To achieve collaborative modeling of visual and semantic information, we have constructed a cross modal fusion module that introduces a language guided mechanism to embed semantic priors in image features, thereby enhancing the understanding and discriminative ability of key regions.As illustrated below, the image feature map $F_v \in R^{H\times W\times D}$is first projected to obtain the Query vectors:

$$Q = F_v W_Q \tag{1}$$

where $W_Q \in R^{D\times D}$

Textual Embeddings as Key/Value: The semantic prototypes are projected to produce Key and Value matrices:

$$K = SW_k \tag{2}$$

$$V = SW_v \tag{3}$$

where $W_k, W_v \in R^{D\times D}$

Modality-Specific Enhancements aligns visual and textual features through dual-path interaction:

$$Attention(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d}})V \tag{4}$$

where $d = \frac{D}{h}$is the dimension per attention head (h is the number of heads)

For multi-head extension,h parallel heads are concatenated and linearly mixed:

$$MultiHead(Q,K,V) = Concat(head_1, head_2, \ldots, head_h)W_o \tag{5}$$

$$head_i = Attention(QW_Q^i, KW_K^i, VW_v^i) \tag{6}$$

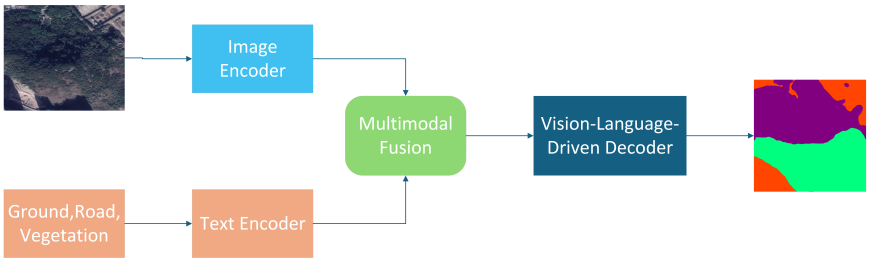

Figure 1. MultiTrans-LC.We extract features from images; Visual embedding is an advanced feature mapping used for image embedding; Visual language information is the combination of image features and text features.

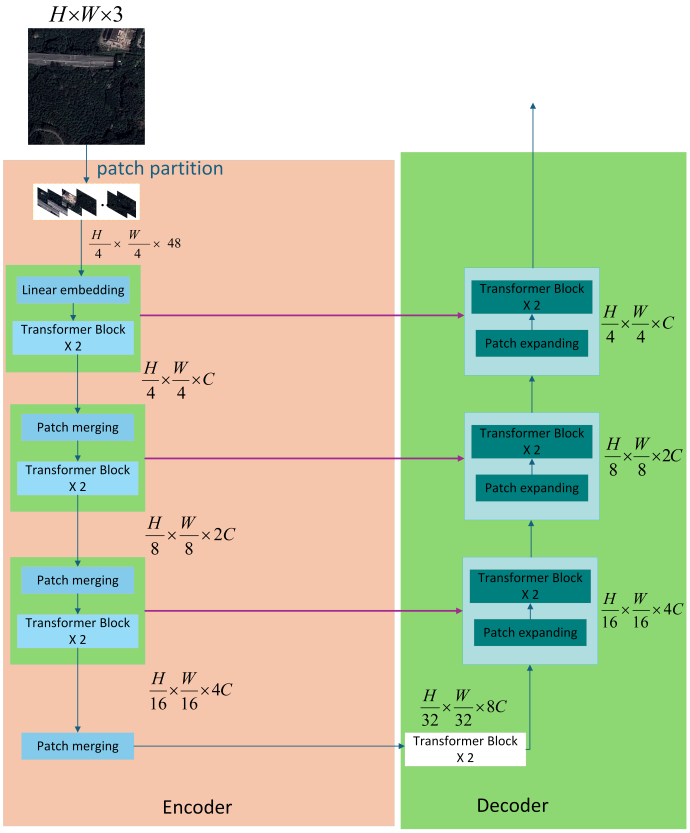

Figure 2. The overall structure of Visual Transformer.

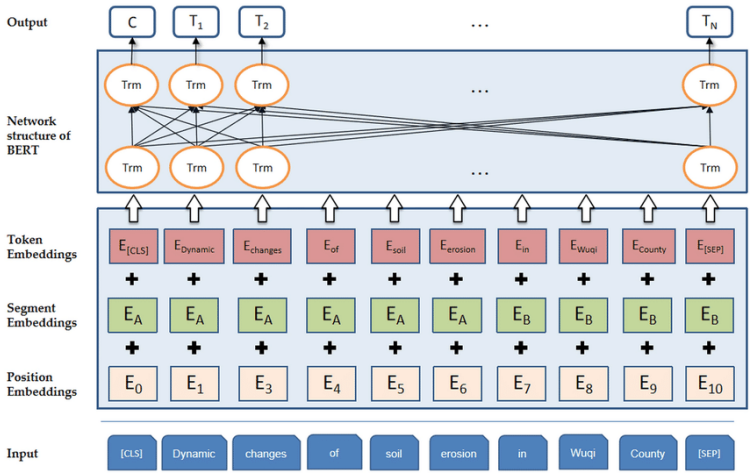

Figure 3. The overall structure of Textual Transformer.

A gating mechanism balances multimodal contributions:

$$\alpha = \sigma(MLP([F_v; S]))(Sigmoidactivation) \tag{7}$$

$$F_{final} = \alpha \cdot F_v + (1 - \alpha) \cdot S \tag{8}$$

Final fusion features:

$$F_{fused} = MultiHead(Q, K, V) \tag{9}$$

The above mechanism enables the network to emphasize regions that are semantically consistent with the given textual description, yielding discriminative features for subsequent decoding.

3.4 Vision-Language-Driven decode

To strengthen MultiTrans-LC's learning ability in the decoding stage, we designed a layer-by-layer multi-level feature fusion structure based on the Transformer block. We transfer feature information from high-level to low-level and use Transformer block to improve the decoding stage representation in each feature fusion level. In the encoding stage of MultiTrans-LC, we fused text semantic information to the image feature map. In the decoding stage of MultiTrans-LC, we performed a multimodal decoder with multi-level feature fusion module on remote sensing images. As a result, we finally obtained the decoding prediction result.(Dong et al., 2024)

3.5 Edge-Aware Loss

In MultiTrans LC, to solve the problem of boundary blurring, to alleviate the problem of unclear recognition of boundary regions in traditional methods, we have introduced an optimization mechanism specifically designed for edge or changing regions. Through multi strategy feature fusion or joint loss functions, the model can improve its perception of detailed structures while maintaining semantic accuracy. The mathematical expression is as follows:

$$L = L_{cls} + \lambda \cdot L_{boundary} \tag{10}$$

L_{cls}Using standard Cross Entropy Loss, optimize the semantic classification results of the model for each pixel:

$$L_{cls} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(p_{i,c}) \tag{11}$$

where N is the total number of pixels, and C is the number of categories.$y_{i,c}$ is the real label of the i-th pixel (one hot encoding).$p_{i,c}$ is the probability that the i-th pixel predicted by the model belongs to category c.

$$L_{boundary} = \frac{1}{M} \sum_{j=1}^{M} ||\check{B}_j - B_j||_1 \tag{12}$$

where M is the number of boundary pixels, $\check{B}_j$ is the predicted boundary probability map of the model, and B_j is the true boundary mask (labeled data). The function of the balance factor λ is to adjust the weights of classification loss and boundary loss, preventing one from dominating the optimization process.In the paper, λ is set to 0.5, and experiments show that this value can balance classification and boundary optimization objectives in most scenarios.

In summary, we propose a powerful visual language driven decoder for land classification tasks. This decoder utilizes visual language and Transformer module to construct a global attention mechanism, which can accurately decode multi-level image information and enhance the feature expression ability in the decoding stage. In addition, we combine the visual language features of the encoding stage with the visual features of the decoding stage to enrich the semantic information of the decoding stage and accurately extract key change features from remote sensing images. Finally, we designed a decoding module architecture that integrates features layer by layer. This structure based on Transformer modules ensures effective transmission of feature information from high to low layers, fully utilizing key features in the decoding process, and thereby improving the performance of remote sensing image change detection.

4. Experiments

4.1 Experimental Settings

Our framework is built using the PyTorch programming environment and enhanced with the mmsegmentation library. The experimental setup is hosted on the Ubuntu system and equipped with NVIDIA GeForce RTX 3090 GPU to accelerate model training. In terms of model optimization, AdamW optimizer is used with a learning rate set to 0.0001 and weight decay parameter set to 0.01. Our loss function is cross entropy, which only calculates the decoder output during training. Throughout the entire experimental phase, we continuously monitored the mIoU metric on the validation set to specify the best performing model for the subsequent final evaluation. This consistent approach makes the evaluation of Multi-Trans LC on datasets powerful and effective.

4.2 Evaluation metrics

To evaluate the performance of our algorithm, in addition to overall accuracy (OA), we mainly use F1 score, kappa coefficient, and joint interaction (IoU) to assess change detection evaluation metrics. In tasks with multiple variables, the higher the accuracy value, the lower the false positive rate of the predicted results, and the higher the recall value, the lower the missed detection rate of the predicted results. F1 score, mF1, kappa, IoU, mIoU, and OA are all comprehensive evaluation indicators. Larger values indicate better predictions. The calculation formula is as follows:

$$F1 = 2\frac{P \cdot R}{P + R} \tag{13}$$

$$mF1 = \frac{1}{N} \sum_{1}^{N} 2\frac{TPn}{TPn + FPn + FNn} \tag{14}$$

$$kappa = \frac{p_o - p_e}{1 - p_e} \tag{15}$$

$$IoU = \frac{TP}{TP + FN + FP} \tag{16}$$

$$mIoU = \frac{1}{N} \sum_{1}^{N} \frac{TPn}{TPn + FPn + FNn} \tag{17}$$

$$OA = \frac{TP + TN}{TP + TN + FN + FP} \tag{18}$$

where TP is the number of true positive predictions, TN is the number of true negative predictions, FP is the number of false positive predictions, and FN is the number of false neg-

ative predictions. p_o is the observed consistency ratio, which is the ratio of the model's predicted results to the actual classification results. p_e is the expected consistency ratio, which is the probability of the model's predicted results being consistent with the actual classification results in a random situation.These metrics provide a comprehensive evaluation of the algorithm's performance in terms of its ability to detect and classify changes in remote sensing images.

4.3 Datasets

To verify the generalization ability of the proposed framework, three open-source datasets were selected for the experiment, and the training and testing sets were divided in an 8:2 ratio to ensure objective and reliable evaluation results. we first selected three datasets, GID15, LoveDA, and RSSCN7, for land classification. These three datasets all have sub meter resolution, and with the advancement of satellite technology, we believe that these high-resolution remote sensing images are more representative. Overall, to ensure fair experimental comparisons, we chose three public datasets (GID15, LoveDA, and RSSCN7 datasets) and separated the training and testing sets. Below is a detailed introduction to these datasets.

GID15: The land cover information is widely distributed, including 150 high-quality and high-resolution Gaofen-2 satellite remote sensing images, covering a geographic area of over 50000 square kilometers, involving more than 60 different cities in China, and the images are clear and of high quality without cloud cover. Has extremely rich diversity in spectrum, texture, and structure, closely resembling the real distribution characteristics of land features

LoveDA: It contains 5987 high-resolution images and 166768 annotated semantic objects from three different cities: Nanjing, Changzhou, and Wuhan. The annotated semantic categories include background, buildings, roads, water bodies, bare soil, forest land, cultivated land, sports fields, etc. The challenges mainly come from multi-scale objects, complex background samples, and inconsistent class distributions.

RSSCN7: Contains 2800 remote sensing images, divided into 7 typical scene categories: grassland, forest, farmland, parking lot, residential area, industrial area, and river and lake. Each category contains 400 images, with a pixel size of 400 × 400 per image.

4.4 Analysis and visualize results

As shown in table 1-3,we extensively evaluated on MultiTrans-LC three datasets-GID15, LoveDA, RSSCN7. MultiTrans LC achieved the best performance on all datasets.

Our experimental results indicate that on the GID15 dataset, MultiTrans-LC outperforms recent classical models in all metrics.MultiTrans-LC leads with 90.7% OA, which is 3.4% and 1.6% higher than DeepLabV3+ and Swin UNet, respectively, indicating its stronger global classification ability.The Kappa value of MultiTrans-LC reaches 0.901, which is much higher than other models, indicating that its classification results are more consistent with real labels, especially in scenarios with uneven class distribution, and it performs more robustly.MultiTrans LC leads with an IoU of 84.3%, showing a significant improvement compared to Swin UNet (82.5%) and DeepLabV3+(80.4%), attributed to its edge aware loss function directly optimizing the boundary region.The optimal mIoU res-

Model	OA (%)	Kappa	mF1 (%)	F1(%)	IoU(%)	mIoU(%)
DeepLabV3+	87.3	0.832	79.2	86.5	80.4	86.6
Swin-UNet	89.1	0.851	81.5	88.6	82.5	86.8
MultiTrans-LC	**90.7**	**0.901**	**83.9**	**90.1**	**84.3**	**88.9**

Table 1. Quantitative Results on the GID15 dataset.

Model	OA (%)	F1(%)	IoU(%)
DeepLabV3+	84.7	85.2	84.3
Swin-UNet	88.2	87.1	86.5
MultiTrans-LC	**90.5**	**88.9**	**87.9**

Table 2. Quantitative Results on the LoveDA dataset.

Model	OA (%)	F1(%)	IoU(%)
DeepLabV3+	85.3	84.3	82.4
Swin-UNet	87.2	88.1	84.5
MultiTrans-LC	**89.6**	**89.7**	**86.9**

Table 3. Quantitative Results on the RSSCN7 dataset.

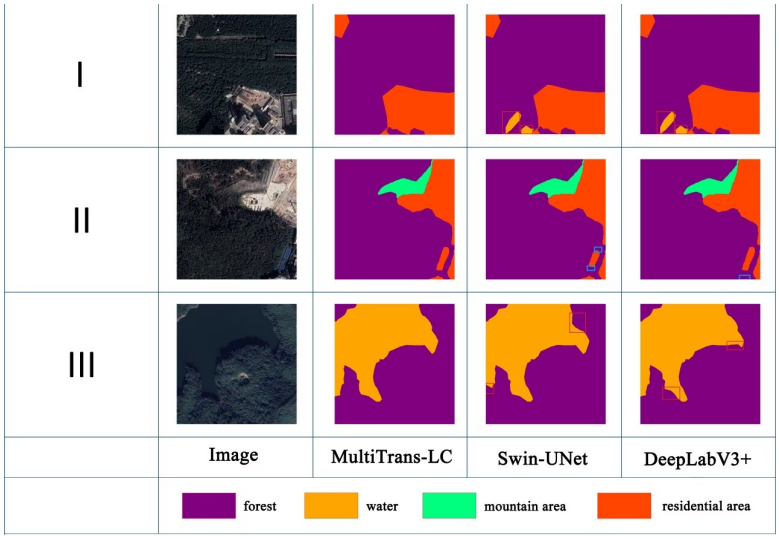

Figure 4. Example of classification result visualization.

ults demonstrate its comprehensive performance advantage in multi class scenarios.

MultiTrans LC's comprehensive leadership on two other datasets validates the effectiveness of its multimodal fusion, providing an efficient solution for land cover classification in complex remote sensing scenarios.

Select three cropped images to visualize the semantic segmentation results of different models on the dataset. As shown in Figure 4.

In group I, The visualization results show that the comparison method exhibits significant misclassification in the boundary region, while the model proposed in this paper can accurately locate the target contour with smoother and more continuous edges.

In group II, Swin UNet and DeepLabv 3+ were almost unable to identify small patch areas in residential areas, but MultiTrans LC could accurately identify and optimize the category segmentation range.

In group III, MultiTrans LC recognizes forests and water bodies, while Swin UNet and DeepLabv 3+can recognize some of their contours, but the edges are rough and not accurately recognized. MultiTrans LC is more effective in identifying the entire area, with smoother and more continuous edges.

Overall, MultiTrans LC outperforms other models in classification performance, improving phenomena such as incomplete classification, unclear boundaries, misclassification, missed classification, and over segmentation, significantly improving recognition accuracy.

5. Conclusion

By integrating visual and textual features, combining cross modal attention mechanism and edge perception loss function, it significantly improves the classification accuracy and boundary clarity in complex scenes. This framework uses visual Transformer to extract global contextual features of remote sensing images, and guides the model to focus on key regions through language description. At the same time, a hierarchical multi-level feature fusion decoder is introduced to optimize multi-scale information expression. Through extensive experimentation on three benchmark datasets (GID-15, LoveDA, and RSSCN7), the model demonstrates superior performance over existing state-of-the-art methods, achieving an overall accuracy of 90.7% and a Kappa coefficient of 0.901 on the GID-15 dataset. Its core innovation lies in the language driven semantic alignment mechanism and joint boundary optimization strategy, effectively solving the problems of small target missed detection and boundary blurring, providing high-precision land cover mapping solutions for applications such as urban planning and disaster assessment. However, three limitations warrant consideration. First, the reliance on predefined text templates restricts adaptability to novel land cover types outside training vocabularies. Second, while the current implementation processes 512×512 pixel tiles efficiently, scaling to continental-scale mapping requires further optimization for ultra-high-resolution imagery (e.g., 30,000×30,000 pixels). Third, performance variations across seasonal conditions indicate the need for temporal adaptation modules to handle phenological changes in vegetation. Future research will focus on three directions:

(1) Using large-scale language models such as GPT-4 (LLM) to generate adaptive text descriptions based on regional geographic knowledge, achieving open vocabulary classification.

(2) Real time processing of edge devices installed on drones through neural structure search and quantification techniques.

(3) By integrating multi temporal Sentinel-2 time series to capture land cover evolution patterns, robustness to seasonal changes can be improved. In addition, extending the framework to 3D city modeling using LiDAR text fusion can open up new applications in smart city planning.

In conclusion, MultiTrans-LC establishes a new paradigm for geospatial artificial intelligence by bridging the gap between visual perception and linguistic semantics. Its technical innovations not only advance the state of remote sensing analysis but also lay the groundwork for human-AI collaborative systems where domain experts can refine results through natural language interactions. As satellite constellations increasingly deliver petabyte-scale Earth observation data, frameworks like MultiTrans-LC will be indispensable for transforming raw pixels into actionable environmental intelligence, ultimately supporting global sustainability goals such as the UN's 2030 Agenda for Sustainable Development.

References

Dong, S., Wang, L., Du, B., Meng, X., 2024. ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208, 53-69. https://www.sciencedirect.com/science/article/pii/S0924271624000042.

Kalyan, K. S., Rajasekharan, A., Sangeetha, S., 2021. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*.

O'shea, K., Nash, R., 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Wang, Y., Yang, L., Liu, X., Yan, P., 2024. An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3+. *Scientific reports*, 14(1), 9716.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., Tay, F. E., Feng, J., Yan, S., 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *Proceedings of the IEEE/CVF international conference on computer vision*, 558–567.

Zheng, D., Dong, W., Hu, H., Chen, X., Wang, Y., 2023. Less is more: Focus attention for efficient detr. *Proceedings of the IEEE/CVF international conference on computer vision*, 6674–6683.