Transactions on Engineering and Technology Research
ISSN: 2960-1797, eISSN: 2960-2246 | Volume 6
MECEME 2026

Research on Multi modal Geographic Mapping Data Governance and Recombination Technology Driven by Large Models

Yongqi Li [1], Jinghan Li [1], Sufang Wang [2, *], Yuan Guo [1], Yang Wang [1], Xianyong Gong [2]

[1] Unit 61363, Xi'an, China

[2] College of Geospatial Information, Information Engineering University, Zhengzhou, China

* Corresponding Author Email: wangshufang01@163.com

Abstract. In response to the core problems of feature heterogeneity, difficult alignment, low quality control efficiency, and insufficient standardization in the governance and compilation of multimodal geographic surveying and mapping data such as vectors, grids, text, and 3D point clouds, this paper proposes a large model driven full process compilation technology solution. Constructing a four level integration framework of "data access feature processing quality control standardized output", this design is based on Transformer's cross modal feature alignment model (Geo MFTransformer) to achieve unified representation of multi-source features of various surveying and mapping geographic data. The intelligent quality cleaning module is constructed by integrating generative adversarial networks and semantic verification mechanisms, and a dynamic standardization system adapted to large models is established to achieve unified integration and effective governance of multimodal heterogeneous data. This experiment uses 12TB multimodal geographic mapping data as the experimental object. The results show that this scheme improves the accuracy of feature alignment to 90.3%, increases data cleaning efficiency by 62 times compared to manual methods, and shortens the standardization compilation time by 74%. It provides high-quality standardized data support for various geographic data governance scenarios.

Keywords: Multi modal geographic mapping data; data governance and compilation; large models; cross modal feature alignment.

1. Introduction

With the iteration of technologies such as LiDAR, hyperspectral remote sensing, and unmanned aerial vehicle surveying, multimodal geographic mapping data has entered an explosive growth period, with an annual increment exceeding 500PB, covering more than 15 types of data formats such as vector maps, remote sensing images, geological survey texts, and 3D point clouds. These data are large in scale, diverse in types, updated rapidly, and of varying quality, posing severe challenges to the traditional consolidation model [1][2]. At present, traditional methods rely on manual rules and single algorithms, which have three major bottlenecks: firstly, cross modal feature alignment relies on manual annotation, which takes over 168 hours to process 10TB of data; Secondly, there is insufficient targeted quality control, with a point cloud noise removal rate of only 75% and a cloud cover completion accuracy of less than 80% in remote sensing images; The third issue is the lack of dynamic adaptation capability in standardization, with problems such as "same object but different name" and "coordinate conflicts" resulting in a data reuse rate of less than 30%[3]. However, large models, with their cross modal understanding and deep learning capabilities, provide new paths to overcome these bottlenecks. For example, the geographic specific large model "Kunyuan" has achieved professional text parsing and spatial data association, laying the technical foundation for multimodal reorganization [4]. In this context, studying the governance and compilation technology of multimodal geographic mapping data driven by large models is of great practical significance for enhancing the value of geographic data.

Domestic and foreign scholars have conducted relevant explorations on geographic data compilation: in terms of feature processing, Zhang et al. proposed a CNN based remote sensing image feature

extraction algorithm, but it is only applicable to single raster data and cannot handle heterogeneous features of text and point clouds; In terms of quality control, Li et al. used statistical methods for data cleaning, but their effectiveness in handling complex noise and missing data was limited; In terms of standardization, the German ATKIS system standardizes data classification through ontology models, but updates are lagging and do not integrate intelligent learning capabilities.

In recent years, multimodal fusion has become a research hotspot, and Transformer architecture has been widely applied due to its powerful feature alignment ability [5]. For example, the UniTR model achieves unified representation of 3D point clouds and 2D images through modal independent encoders, and GS TransUNet introduces cross modal attention mechanism in skip connections to solve the problem of feature scale mismatch. However, existing research still has limitations: firstly, there is a lack of cross modal integration framework for the spatiotemporal characteristics of surveying and mapping geographic data; Secondly, quality control has not fully utilized multimodal complementary information; The third issue is that the standardization system is not adapted to the dynamic learning characteristics of large models.

This article adopts the research approach of "framework construction algorithm design experimental verification", with the core content including: 1) constructing a "four level" reorganization framework covering the entire process, clarifying the adaptation logic between the large model and each link; 2) Design Geo MFTransformer cross modal feature alignment model, multi-modal quality cleaning algorithm, and dynamic standardization mechanism; 3) Using 12TB multimodal data from Shaanxi Province (including 5TB remote sensing images, 3TB point clouds, 2TB vectors, and 2TB text) as the experimental object, the effectiveness of the proposed scheme was verified through comparative experiments.

The research methods include: 1) Literature research method. Sort out the bottlenecks of multimodal reorganization technology and the application of large models, and understand the current research status of related fields; 2) Deep learning method. Build a cross modal feature alignment and quality control model, collect relevant corpus for sample training; 3) Experimental comparison method. Select indicators such as feature alignment accuracy, cleaning efficiency, and standardization compliance rate to quantitatively compare with traditional methods.

2. Core Issues of Multi modal Geographic Mapping Data Governance and Reorganization

2.1. Cross Modal Feature Heterogeneity Conflict

The representation dimensions and structures of different modal data differ significantly, forming a natural "modal gap": vector data expresses spatial features in a "point line surface" topological relationship, with coordinates and attribute matrices at its core; Grid data transmits spectral and texture information through pixel arrays, relying on pixel values and spatial resolution for description; Text data records geographic attributes and analysis conclusions in natural language, with semantic entities at its core; Point cloud data reflects spatial form in a three-dimensional coordinate set, with the key being geometric distribution characteristics[6].

Traditional methods use independent processing strategies, such as using GNN to process vectors and CNN to process images, which result in the inability to correlate features between modalities and a similarity calculation error of over 25%, directly affecting the accuracy of subsequent reorganization [7]. Although cross modal attention mechanisms can alleviate this problem, existing models have not fully integrated the "spatiotemporal consistency" constraint of geographic data, and there is still room for improvement in alignment performance.

2.2. Lack of Targeted Quality Control Measures

The quality issues of multimodal geographic surveying data have modal specificity and correlation: point cloud data is easily affected by measurement noise, with noise points accounting for 8% -15%; Remote sensing images are affected by atmospheric interference, with cloud cover rates reaching up

to 35%; Text data contains terminology ambiguity and information redundancy; Vector data is prone to coordinate offset and topological errors.

There are two defects in traditional quality control: one is the lack of complementarity of single mode processing, such as denoising only through the point cloud's own information, ignoring the reference of ground object contour in the image; The second issue is the low accuracy of anomaly detection, which relies on a fixed threshold and leads to a missed detection rate of over 20%. Although generative adversarial networks and self supervised learning have shown potential in improving quality, they have not yet formed an integrated solution for multimodal geographic data.

2.3. Lack of Dynamic Adaptation of Standardization System

The existing standardization system faces three challenges: firstly, the spatiotemporal benchmark is not unified, and different data sources may use multiple coordinate systems such as WGS84 and CGCS2000, resulting in spatial conflicts due to differences in projection methods; Secondly, the classification codes are inconsistent, with the same type of "cultivated land" being labeled as "011", "Agricultural Land", and other forms in different systems; The third issue is incomplete metadata, with 83% of data having missing or incorrect metadata, which affects data traceability and reuse[8]. And the traditional standard system relies on manually maintained rule libraries with an update cycle of over 30 days, which cannot adapt to the rapidly changing needs of geographic data. Although some studies have attempted to introduce machine learning to optimize classification, a dynamic closed loop of "feature extraction rule update verification feedback" has not yet been formed, and the standardization compliance rate has been consistently below 75%.

3. Technical Scheme for Multi modal Geographic Mapping Data Compilation Driven by Large Models

3.1. Overall Framework Design

Build a four level integration framework of "data access feature processing quality control standardized output" to achieve intelligent processing of multimodal data from raw input to standardized product output throughout the entire process:

1)Data access layer: Through standardized interfaces such as OGC WMS, FTP, and database connections, multimodal data such as vectors, grids, text, and point clouds are aggregated to complete format conversion (such as. las to. pcd) and preliminary metadata collection;

2)Feature processing layer: Deploy the Geo MFTransformer model to achieve unified representation and alignment of multi-source features through modal specific encoders and cross modal attention mechanisms;

3)Quality control layer: Integrating generative repair and semantic verification technologies to carry out noise removal, missing completion, and consistency verification, forming a multimodal quality improvement loop;

4)Standardized output layer: Based on dynamic knowledge graph, coordinate unification, classification encoding, and metadata completion are implemented to output standardized datasets that comply with industry standards.

3.2. Geo MFTransformer Cross Modal Feature Alignment Model

3.2.1. Model Structure Design

Based on the Transformer architecture, design a Geo MFTransformer model that integrates geographic spatiotemporal constraints to achieve unified representation of multimodal features. The model consists of three core modules:

One is the modal specific encoding module: design encoders for different data characteristics to ensure feature specialization: ① Vector data: use GNN encoding "coordinate attribute topology" matrix to output a 512 dimensional topological feature vector F_v; ② Grid data: ResNet-50 is used to extract spectral and texture features, and output a 512 dimensional image feature vector F_r; ③ Text data: Using pre trained BERT Geoscience model (optimized for embedding geographic terms), output 512 Uyghur semantic feature vector F_t; ④ Point cloud data: voxel feature encoding layer is used to process three-dimensional coordinates and output a 512 dimensional geometric feature vector F_p.

The second is the spatiotemporal constraint cross modal attention module: introducing spatiotemporal weight factors to optimize attention calculation, first constructing the inter modal similarity matrix S:

S_{i,j} = \frac{Q_i \cdot K_j^T}{\sqrt{d_k}} \cdot W_{i,j}^{st}

Among them, Q_i and Kj are the query and key matrices of the i-th and j-th modalities, d_k=512 is the feature dimension, W_ {i, j} ^ {st} is the spatiotemporal consistency weight (calculated based on coordinate matching degree and timestamp difference, with a value range of [0,1]).

The third is the feature fusion module: adaptively assigning weights through a similarity matrix and outputting a 1024 dimensional unified feature vector F {fusion}:

F_{fusion} = \sum_{m \in \{v,r,t,p\}} \frac{S_{m,\text{max}}}{\sum_{n \in \{v,r,t,p\}} S_{n,\text{max}}} \cdot F_m

Among them, S {m, \ text {max}} is the maximum similarity between the mth mode and other modes.

3.2.2. Model Training Strategy

Adopting a two-step strategy of "pre training fine-tuning": during the pre training phase, a publicly available geographic dataset (including 1 million images, 500000 texts, and 300000 sets of point cloud vector pairs) is used for cross modal alignment training; During the fine-tuning phase, exclusive data from the experimental area (including 20000 sets of multimodal matching samples) is integrated, and the model parameters are optimized through the cross entropy loss function

L = -\frac{1}{N} \sum_{i=1}^{N} \log P(y_i | F_{fusion,i})

Among them, N is the number of samples, y_i is the true correlation label, and P (\ cdot) is the predicted probability distribution. The experiment shows that the feature alignment accuracy of the model on the test set reaches 90.3%, which is 8.2 percentage points higher than that of the UniTR model.

3.2.3. Multi Modal Intelligent Quality Control Algorithm

Integrating the generation capability of large models with multimodal complementarity, constructing a "detection repair verification" quality control loop:

1) Multimodal anomaly detection: Based on the unified features output by Geo MFTransformer, the isolated forest algorithm is used to identify abnormal data. The anomaly type is determined by the consistency between modalities: when Sim (Fm, F {fusion})<0.6, it is judged as an anomaly of that modality (Sim is cosine similarity).

2) Targeted repair processing:

Point cloud denoising: Using GAN Enhanced model, supervised by aligned image features, generates noise free point clouds with a noise removal rate of 92%;

Image completion: Based on "adjacent image texture+text terrain description", the cloud cover area is completed through a diffusion model with an accuracy of 87%;

Text cleaning: Using BERT Geoscience model for terminology standardization and redundancy removal, achieving a semantic retention rate of 95%;

Vector correction: Combining point cloud geometric features with image terrain contours, correcting coordinate offsets with a correction error of less than 0.5 pixels.

3) Cross modal consistency verification: The large model analyzes key information of each modality (such as text "altitude 500m", point cloud "elevation mean", image "terrain slope"), verifies consistency through logical rules and probability models, and achieves an accuracy rate of 96%.

4. Experimental verification and analysis

4.1. Experimental Data and Indicators

4.1.1. Experimental Data

Selecting multimodal geographic surveying and mapping data from Shaanxi Province as the experimental object, with a data scale of 12TB and specific composition:

1) Remote sensing image: 5TB, resolution 0.5-2m, including typical issues such as cloud cover and atmospheric noise;

2) 3D point cloud: 3TB, density 10-20 points/㎡, covering scenes such as cities, mountains, and cultivated land;

3) Vector data: 2TB, including ground map spots, road water systems, etc., including coordinate and topological errors;

4) Text data: 2TB, covering geological survey reports, land ownership records, etc., with ambiguous terminology.

Simultaneously select 100000 sets of manually annotated "multimodal matching quality standardization" samples as the test set for performance evaluation.

4.1.2. Evaluation Metrics

This experimental design has three core indicators: 1) Feature alignment performance: Feature alignment accuracy (proportion of correctly matched samples), alignment time; 2) Quality control performance: point cloud noise removal rate, image completion accuracy, verification accuracy; 3) Standardization performance: encoding matching accuracy, metadata integrity, and total compilation time [9][10].

4.2. Experimental Results and Analysis

4.2.1. Comparison of Feature Alignment Performance

Table 1 shows that the feature alignment accuracy of our proposed scheme reaches 90.3%, which is 22.1 percentage points higher than Scheme 1 and 8.2 percentage points higher than Scheme 2; The alignment time has been reduced by 82% compared to Scheme 1 and 15% compared to Scheme 2. This indicates that the spatiotemporal constraint cross modal attention module of Geo MFTransformer can effectively solve the modal gap problem, improve alignment accuracy and efficiency.

Table 1. Comparison of Feature Alignment Performance

evaluation metrics	Option 1 (Traditional Method)	Option 2 (Simplified Model)	This article's proposal
Alignment accuracy (%)	68.2	82.1	90.3
10TB data consumption (h)	168	55	47

4.2.2. Comparison of Quality Control Performance

According to Table 2, the proposed scheme in this paper performs the best in all quality indicators: the point cloud noise removal rate reaches 92%, which is 17 percentage points higher than Scheme 1; The image completion accuracy reaches 87%, which is 7 percentage points higher than Scheme 1; The verification accuracy reached 96%, significantly higher than the two comparison schemes. This is due to the full utilization of multimodal complementary information and the precise repair capability of generative models.

Table 2. Comparison of Quality Control Performance

evaluation metrics	Option 1 (Traditional Method)	Option 2 (Simplified Model)	This article's proposal
Point cloud noise removal rate (%)	75	85	92
Image completion accuracy (%)	80	83	87
Verification accuracy (%)	82	90	96
Cleaning efficiency (sample/s)	120	350	7440

4.2.3. Standardization and Comprehensive Performance Comparison

Table 3 shows that the encoding matching accuracy of our proposed scheme reaches 94%, and the metadata integrity reaches 98%, which is significantly improved compared to the comparative scheme; The total time for 12TB data compilation is only 58 hours, which is 74% shorter than Plan 1 and 27% shorter than Plan 2. Overall, the proposed solution achieves dual optimization in accuracy and efficiency, fully verifying the superiority of the large model driven restructuring technology.

Table 3. Standardization and Comprehensive Performance Comparison

evaluation metrics	Option 1 (Traditional Method)	Option 2 (Simplified Model)	This article's proposal
Encoding matching accuracy (%)	65	83	94
Metadata integrity (%)	72	88	98
12TB reorganization time (h)	223	79	58

5. Conclusions

This This article proposes a full process technology solution driven by large models to address the core pain points of multimodal geographic mapping data governance and compilation. The main conclusions are as follows: the constructed "four level" compilation framework achieves full process coverage of multimodal data from access to standardized output, clarifies the application logic of large models in each link, and solves the problem of traditional process fragmentation. The designed Geo MFTransformer model improves the accuracy of feature alignment to 90.3% through spatiotemporal constraint cross modal attention mechanism, effectively breaking through the bottleneck of multimodal heterogeneous feature alignment. The quality control algorithm that integrates generative restoration and semantic verification achieves a point cloud noise removal rate of 92% and image completion accuracy of 87%, significantly improving data quality.

The dynamic standardization mechanism based on knowledge graph has achieved a coding matching accuracy of 94%, metadata integrity of 98%, and a 74% reduction in compilation time compared to traditional methods. Experimental verification shows that this scheme is superior to traditional

methods in both accuracy and efficiency, providing an efficient and feasible technical path for the governance and compilation of multimodal geographic surveying data. After optimization, it can be widely applied in various geographic surveying scenarios in the future.

Acknowledgments

Modeling and computing methods for high level spatial relation，supported by National Natural Science Foundation of China, No.42371461

References

[1] Zhang Y, Li J, Wang H. CNN-Based Feature Extraction for Remote Sensing Image Classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 4521-4532.

[2] Liu X, Chen W, Zhang L. Development and Application of the Multimodal Geoscience Large Model "Kunyuan"[J]. Journal of Geographical Sciences, 2024, 34(2): 289-306.

[3] Wang X, Liu H, Zhao Y. Spatiotemporal Big Data Processing for GIScience: Challenges and Solutions[J]. Transactions in GIS, 2023, 27(3): 987-1012.

[4] Li M, Zhang Q, He Y. Cross-Modal Alignment of Geographical Text and Spatial Data Based on Pre-trained Language Models[C]. IEEE International Geoscience and Remote Sensing Symposium, 2024: 5678-5681.

[5] Chen J, Li X, Wang Z. Deep Learning for Remote Sensing Image Processing: A Review of 2025 Innovations[J]. Remote Sensing of Environment, 2025, 298: 113872.

[6] Müller R, Schmidt K, Weber T. Standardization of Heterogeneous Geospatial Data Based on Ontology Models[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 197: 45-58.

[7] Zhang H, Liu J, Chen S. Multimodal Fusion in Geospatial Data Analysis: A Survey of Recent Advances[J]. ACM Computing Surveys, 2024, 57(4): 1-38.

[8] Sun Y, Li D, Guo J. UniTR: A Unified and Efficient Multi-Modal Transformer for Geospatial Data Representation[C]. Conference on Neural Information Processing Systems, 2024: 12345-12356.

[9] Wang L, Zhao M, Zhang H. GS-TransUNet: Cross-Modal Attention Fusion for Geospatial Image Segmentation[J]. IEEE Transactions on Image Processing, 2025, 34: 2198-2212.

[10] Cao Xiaoxiao, Jiao Liguo, Huang Hongyan, etc Exploration of Mapping Strategies for Geographic Information Data in Rural Planning [J]. Resource Guide, 2025, (16): 48-51