# Research on the Method of Using Deep Learning to Improve the Accuracy of Complex and Mature Spatial 3D Remote Sensing Reconstruction

**Shiman Wang**
College of Geography and Planning, Chengdu University of Technology, Chengdu 610059, Sichuan, China
Corresponding author: Shiman Wang (e-mail: wangshiman77899@163.com)

**ABSTRACT** Driven by the construction of “Digital China” and advances in spatial information technology, three-dimensional remote sensing reconstruction of complex environments has become increasingly important for digital twin applications and intelligent spatial analysis. In particular, the integration of electromagnetic remote sensing modalities, including synthetic aperture radar and LiDAR, provides complementary information for high-precision environmental perception and large-scale scene reconstruction. However, complex terrains generally suffer from variable illumination, severe occlusion, heterogeneous land-cover distribution, and sensor noise interference, resulting in insufficient reconstruction accuracy, poor detail preservation, geometric distortion, and limited robustness in conventional methods. To address these challenges, this paper proposes a deep learning-based approach for improving the accuracy of complex spatial 3D remote sensing reconstruction by establishing a unified framework that integrates multimodal data fusion with adaptive noise suppression and collaborative geometric correction. The proposed method enhances the complementary utilization of heterogeneous sensing information and optimizes reconstruction precision in largescale complex scenes, providing effective technical support for intelligent spatial perception and electromagnetic remote sensing applications.

**INDEX TERMS** deep learning, 3D remote sensing reconstruction, multimodal data fusion, electromagnetic remote sensing, synthetic aperture radar

## I. INTRODUCTION

### A. RESEARCH BACKGROUND

COMPLEX mature space refers to a 3D spatial region with complex terrain, diverse ground objects, dense occlusion, and high integration of natural environment and human activities, including mountainous hills, dense urban areas, wetland ecological zones, mining subsidence areas, textile industrial parks, etc. WThrough the deep integration of remote sensing, computer vision, and geographic information science, 3D reconstruction has become a pivotal means for constructing digital twins, extending its utility from smart cities and ecological protection to the precise spatial mapping and structural monitoring of specialized textile industrial zones. This technology plays an irreplaceable role in optimizing the spatial layout of high-capacity textile manufacturing hubs and ensuring the environmental safety of their surrounding resource landscapes [1,2]. As the most frequently interacting area between nature and human activities, complex mature spaces cover various types such as mountains and hills, densely populated urban areas, wetland ecological areas, and mining subsidence areas. The high-precision reconstruction of their three-dimensional structures is of great significance for regional planning, environmental monitoring, disaster warning, and other work. According to the "2023 Report on the Development of Remote Sensing Satellite Applications in China", the annual growth rate of China’s demand for 3D reconstruction of complex and mature spaces is over 35%. However, more than 60% of actual projects cannot meet application requirements due to insufficient reconstruction accuracy and serious loss of details[3-4].

Traditional 3D remote sensing reconstruction methods, such as motion recovery structure SfM and multi view stereo matching MVS, mainly rely on manually designed feature extraction operators (SIFT, SURF, ORB), which have poor adaptability to complex scenes and perform poorly in noise interference, occlusion areas, and reconstruction of small structures. In recent years, deep learning technology has made breakthrough progress in the field of computer vision with its powerful feature learning and nonlinear fitting capabilities, providing a new technological path for improving the accuracy of 3D remote sensing reconstruction

[5]. However, existing deep learning based reconstruction methods still have significant limitations: the effectiveness of multimodal data fusion is insufficient, and the complementary advantages of different data sources have not been fully utilized; The feature extraction network has limited ability to represent the edges and occluded areas of complex scenes; Noise suppression and geometric correction are mostly independent modules, lacking deep integration with deep learning; The personalized optimization design for complex mature spaces is insufficient, and the robustness and generalization ability need to be improved [6-8].

### *B. CURRENT RESEARCH STATUS AT HOME AND ABROAD*

Multimodal data fusion is a key means to improve the accuracy of complex scene reconstruction. Foreign scholars have conducted extensive research on data registration, feature fusion, decision fusion, and other aspects. In terms of feature fusion, scholars have proposed fusion methods based on attention mechanism and Feature Pyramid Network (FPN) to enhance complementary feature extraction of multi-source data [9]; A deep learning based registration and feature fusion model is proposed for the fusion of LiDAR point clouds and optical images, such as PointFusion, which effectively fuses multimodal features by projecting point cloud features into the image space [10]; In terms of the fusion of SAR data and optical images, SAR speckle noise is removed through deep learning while preserving the texture and structural information of both types of data [11].

In summary, the existing research has not yet formed a complete technical system of "multimodal data deep fusion deep learning feature accurate extraction adaptive noise suppression geometric accuracy collaborative correction full scene accurate evaluation", and there are still many challenges to improve the accuracy of 3D remote sensing reconstruction in complex and mature spaces. Therefore, this article aims to construct a highprecision 3D remote sensing reconstruction method that adapts to complex environments, filling the research gap regarding the digital twin modeling of textile industrial land use and the spatial analysis of large-scale textile manufacturing clusters.

## II. RELATED THEORETICAL AND TECHNICAL FOUNDATIONS

High resolution optical images have rich texture and spectral information, which can clearly reflect the appearance and spectral characteristics of land features. Its advantage lies in its strong ability to recognize land features, rich texture details, and ease of feature extraction and matching; The disadvantage is that it is easily affected by lighting conditions, cloud cover, and atmospheric disturbances, resulting in poor imaging performance in shaded and water areas, and a lack of direct depth information [12].

LiDAR (Light Detection and Ranging) point cloud data measures the three-dimensional coordinates and reflection intensity of ground objects through laser pulses, providing high-precision and high-resolution depth information. Its advantage lies in the ability to directly obtain the three-dimensional coordinates of land features, without being affected by lighting conditions, and with high measurement accuracy for terrain, topography, and building height; The disadvantage lies in the lack of texture information, uneven point cloud density, susceptibility to occlusion, and high data storage and processing costs.

SAR (Synthetic Aperture Radar) data, through microwave imaging, has the ability to work 24/7, penetrate obstacles such as clouds and vegetation, and obtain information such as backscattering coefficients of ground objects. Its advantage lies in its ability to distinguish water bodies, vegetation, and buildings without being affected by light and weather; The disadvantage lies in the presence of coherent speckle noise, relatively low spatial resolution, and lack of spectral information [13].

GIS geographic vector data includes administrative divisions, roads, water systems, terrain contour lines, etc., with accurate geographic coordinates and semantic information. Its advantage lies in providing geographic constraint information to assist in multi-source data registration and geometric correction; The disadvantage lies in the long data update cycle and poor adaptability to dynamically changing land features [14].

## III. DESIGN OF MULTI MODAL REMOTE SENSING DATA FUSION MODULE FOR COMPLEX MATURE SPACE

### *A. DATA PREPROCESSING*

Multi modal remote sensing data preprocessing is a prerequisite for ensuring fusion quality and reconstruction accuracy, including data format conversion, geometric registration, radiometric correction, denoising processing, and other steps.

The denoising process in the data preprocessing stage mainly targets the obvious noise in the original data, laying the foundation for subsequent fusion and feature extraction. Optical images use median filtering and Gaussian filtering to remove Gaussian noise and salt and pepper noise; LiDAR point clouds use statistical filtering and radius filtering to remove isolated noise points; SAR data uses Lee filtering and Gamma filtering to suppress coherent speckle noise [15].

### *B. MULTIMODAL DATA FUSION FRAMEWORK DESIGN*

#### 1) Single Modal Feature Extraction

Design exclusive feature extraction branches based on the characteristics of different modal data to extract the core features of each modal data:

Optical image feature extraction: Using an improved ResNet-50 network and introducing attention mechanism, texture features, edge features, and spectral features of the image are extracted. The network consists of 5 convolutional stages, each consisting of convolutional layers, batch normalization layers, ReLU activation functions, and pooling layers; After each convolution stage, a channel attention

module is added to enhance the weights of useful feature channels and improve the specificity of feature extraction [16].

SAR data feature extraction: U-Net network is used to extract texture features and backscatter features of SAR data. U-Net encoder extracts multi-scale features, decoder restores feature map dimensions, and skip connections preserve edge detail information; Add a spatial attention module between the encoder and decoder to focus on important regions in SAR images and suppress noise interference [17].

GIS vector data feature extraction: GIS vector data is converted into raster masks, and $1 \times 1$ convolution kernels are used to extract geographic constraint features, such as the location and shape features of roads, water systems, administrative divisions, etc., providing geographic references for subsequent fusion [18]. GIS vector data is converted into raster masks. $3\times3$ and $5\times5$ multi-scale convolutions are used to extract spatial morphological, boundary and connected domain features. Position encoding is added to retain geographic coordinate information.

#### 2) Multi Scale Feature Fusion

Top down path: Upsampling the high-level feature maps of each modality data to the same dimension as the low-level feature maps;

Horizontal connection: Adding elements between high-level feature maps and low-level feature maps of corresponding scales to fuse high-level semantic features and low-level detail features;

Fusion feature output: Optimize the fused feature map through convolution operation to generate the final multi-scale fusion features [19,20].

#### 3) Fusion of Attention Mechanisms

On the basis of multi-scale feature fusion, a cross modal attention mechanism is introduced to enhance the complementarity of multimodal features and suppress redundant information. The core of cross modal attention mechanism is to calculate the correlation between different modal features and adaptively assign different weights to different modal features:

1. Feature correlation calculation: Calculate the cosine similarity between the features of optical images, LiDAR point clouds, SAR data, and GIS vector data, and quantify the correlation between different modal features;

2. Attention weight generation: Based on feature correlation, the attention weights of each modal feature are generated through the softmax function. The higher the correlation, the greater the weight of the modal feature;

3. Weighted feature fusion: Multiply each modal feature with its corresponding attention weight, and then add the elements to obtain the final cross modal fusion feature.

First, all heterogeneous features (LiDAR, SAR, optical, GIS) are mapped into a unified 128-dimensional common feature space via $1\times1$ convolution to eliminate modality gaps. Then feature correlation is calculated in the same dimension. The correlation is used to generate adaptive attention weights for weighted fusion.

#### 4) Fusion Quality Assessment

To verify the effectiveness of multimodal data fusion, design fusion quality evaluation indicators, including information entropy, standard deviation, edge preservation index, mutual information, etc.:

Information entropy: reflects the richness of information in the fused features. The larger the information entropy, the richer the information contained in the fused features;

Standard deviation: reflects the degree of dispersion of the grayscale values of the fused features. The larger the standard deviation, the higher the contrast of the fused features;

Edge preservation index: reflects the degree to which the fused features preserve the edge information of the original data. The closer the index is to 1, the better the edge preservation effect;

Mutual information: reflects the degree of information sharing between different modal features. The larger the mutual information, the better the fusion effect of multimodal features.

## IV. IMPROVED TRANSFORMER U-NET NETWORK CONSTRUCTION

### *A. OVERALL NETWORK STRUCTURE DESIGN*

Design an improved Transformer U-Net network to address issues such as blurred edges, numerous small structures, and severe occlusion of complex and mature spatial features.

### *B. ENCODER DESIGN*

The core function of the encoder is to extract deep semantic features and global features from multimodal fusion features. It consists of four encoding blocks, each of which includes a Transformer module, convolutional layer, batch normalization layer, ReLU activation function, and max pooling layer

Transformer module: adopting a simplified Transformer encoder structure, including multi head self attention mechanism and feedforward neural network, to capture the long-distance dependency relationship and global features of fused features;

Convolutional layer: using $3 \times 3$ convolutional kernels to extract local features and supplement the local feature extraction capability of the Transformer module;

Batch normalization layer: accelerates model training and improves model stability;

ReLU activation function: Introducing nonlinearity to enhance model fitting ability;

Maximum pooling layer: using $2 \times 2$ pooling kernels with a stride of 2, reducing the dimensionality of feature maps, and improving the computational efficiency and robustness of the model.

To solve the problem of gradient vanishing in deep feature extraction, residual connections are added between each

encoding block. The output of the previous encoding block is directly added to the output of the current encoding block to enhance feature transfer, improve model training stability, and enhance feature extraction capabilities.

### C. DECODER DESIGN

The core function of the decoder is to restore the dimensionality of the feature map, fuse deep semantic features with shallow detail features, and generate a dense feature map. It consists of four decoding blocks, each of which includes an upsampling layer, a concatenation layer, a multi-scale feature fusion unit, and a convolutional layer

Upsampling layer: using transpose convolution operation to enlarge the feature map dimension by 2 times and restore the spatial resolution of the feature map;

Splicing layer: Splicing the upsampled feature map with the feature map of the corresponding stage of the encoder, fusing deep semantic features with shallow detail features;

Multi scale feature fusion unit: introducing three convolution kernels of different scales ($1 \times 1$, $3 \times 3$, $5 \times 5$), extracting multi-scale features of concatenated features, and then fusing multi-scale features through attention mechanism to enhance the adaptability of the model to different scales of land cover;

Convolutional layer: Using $3 \times 3$ convolutional kernels to optimize the fused feature map and improve the consistency of feature representation.

### D. EDGE ENHANCEMENT MODULE

To address the issues of blurred edges and loss of details in complex and mature spatial terrain features, an edge enhancement module is added to the decoder output to enhance the extraction and preservation of terrain edge features. The edge enhancement module adopts an improved Canny edge detection algorithm combined with convolutional neural networks:

1. Use the improved Canny algorithm to extract edge information of multimodal fusion features and generate edge masks;

2. Multiply the edge mask with the feature map output by the decoder to enhance the edge features;

3. Optimize the edge enhanced feature map through a 1x1 convolutional layer to ensure consistency and continuity of features.

Coarse edges are predicted from deep network features, then refined by a learnable edge-guided filter, independent of raw noisy images.

### E. OUTPUT LAYER DESIGN

The output layer uses a 1x1 convolution kernel to map the feature map output by the decoder to a depth map of the same size as the input data. Each pixel value in the depth map represents the depth information of the corresponding ground point, providing a basis for subsequent dense point cloud generation.

### F. NETWORK TRAINING STRATEGY

#### 1) Loss Function Design

Design a hybrid loss function for complex and mature 3D reconstruction, including depth loss, edge loss, and smoothing loss, to comprehensively optimize model performance

Depth loss: Using the L1 loss function, calculate the average absolute error between the predicted depth map and the true depth map of the model, and optimize the accuracy of depth prediction;

Edge loss: Based on edge masks, calculate the L2 loss between the predicted depth map and the true depth map in the edge region, and enhance the accuracy of edge feature restoration;

Smooth loss: Using gradient loss function, calculate the gradient norm of the predicted depth map, constrain the smoothness of the depth map, and reduce geometric deformation.

#### 2) Optimizer Selection and Hyperparameter Setting

Adam optimizer is chosen as the optimizer for model training, which combines momentum gradient descent and adaptive learning rate strategy, with fast convergence speed and strong stability. The initial learning rate is set to 0.001, and a learning rate decay strategy is adopted. The learning rate decays to 0.9 every 10 epochs; Set batch size to 8; Set the number of training epochs to 100; Adopting an early stopping strategy, when the validation set loss does not decrease for 10 consecutive epochs, the training is stopped to avoid overfitting of the model.

Early stopping strategy: training stops when the validation loss does not decrease for 25 consecutive epochs.

#### 3) Data Augmentation

To improve the generalization ability and robustness of the model, data augmentation techniques are used to expand the training dataset, including:

Geometric transformation: Randomly rotate ($-10^\circ \sim 10^\circ$), translate ($\pm$ 5 pixels), and scale (0.8–1.2 times) multimodal fusion features;

Lighting transformation: Randomly adjust the brightness (0.7–1.3 times) and contrast (0.7–1.3 times) of optical image features;

Noise addition: Randomly add Gaussian noise (variance 0.01–0.05) and salt and pepper noise (density 0.01–0.03) to simulate the noise interference in actual scenes;

Occlusion simulation: Randomly add rectangular occlusion areas (5%–10% area) to simulate the occlusion of ground objects in actual scenes.

Occlusion simulation: Use building shadow occlusion, high-rise parallax occlusion, and vegetation canopy occlusion instead of random rectangles, conforming to remote sensing imaging physics. The Performance Comparison of Different Models is shown in Table 1:

**TABLE 1.** Performance Comparison of Different Models (ISPRS Benchmark Test Set)

| Model | RMSE (m) | MAE (m) | SSIM |
|---|---|---|---|
| Basic U-Net | 1.86 | 1.32 | 0.72 |
| Transformer | 1.65 | 1.18 | 0.78 |
| MVSNet | 1.52 | 1.05 | 0.81 |
| Improved Transformer-U-Net | 0.94 | 0.63 | 0.92 |

## V. ADAPTIVE NOISE SUPPRESSION AND GEOMETRIC CORRECTION MODEL

### A. DESIGN OF ADAPTIVE NOISE SUPPRESSION MODEL

Multiple sources of remote sensing data in complex and mature spaces are subject to various noise interferences, such as sensor noise, atmospheric disturbances, and coherent speckle noise, which directly affect the reconstruction accuracy.

Design an adaptive noise suppression submodule for different types of noise, and dynamically select the corresponding suppression strategy based on the noise type recognition results

Salt and pepper noise suppression submodule: Combining median filtering and convolutional neural networks, the salt and pepper noise is initially removed through median filtering, and then the image details are optimized through CNN to avoid loss of details;

Coherent speckle noise suppression submodule: using an attention based U-Net network to enhance the texture features of SAR data, suppress coherent speckle noise, and preserve ground structure information;

Atmospheric disturbance noise suppression submodule: Based on the atmospheric correction model and CNN, dynamically adjust the brightness and contrast of the image to eliminate the impact of atmospheric disturbance on image quality.

### B. EVALUATION OF NOISE SUPPRESSION EFFECT

Using peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and noise suppression ratio (NSR) as evaluation indicators for noise suppression effectiveness:

PSNR: Reflecting the clarity of the image after noise suppression, the higher the PSNR, the better the image quality;

SSIM: Reflects the structural similarity between the noise suppressed image and the original noise free image. The closer the SSIM is to 1, the better the structural preservation effect;

NSR: Reflecting the strength of noise suppression, the larger the NSR, the more significant the noise suppression effect.

Quantitatively evaluate the noise suppression effect through the above indicators to ensure that the feature map after noise suppression not only eliminates noise interference, but also preserves the structure and detail information of the ground objects.

### C. CONSTRUCTION OF GEOMETRIC CORRECTION MODEL

Using root mean square error (RMSE), mean absolute error (MAE), and relative error (RE) as evaluation metrics for geometric correction effectiveness:

RMSE: Reflects the overall error of the three-dimensional coordinates of the ground cover points after correction. The smaller the RMSE, the higher the correction accuracy;

MAE: Reflects the average error of the three-dimensional coordinates of the ground objects after correction. The smaller the MAE, the more stable the correction effect;

RE: Reflects the relative error of the three-dimensional coordinates of the ground cover points after correction. The smaller the RE, the better the consistency of the correction.

Quantitatively evaluate the geometric correction effect through the above indicators to ensure that the geometric accuracy of the reconstructed results after correction meets the application requirements of complex and mature space 3D reconstruction.

### D. COLLABORATIVE OPTIMIZATION OF NOISE SUPPRESSION AND GEOMETRIC CORRECTION

Noise suppression and geometric correction are not independent processes. Noise interference can affect the accuracy of geometric deformation detection, and geometric deformation may also lead to uneven noise distribution. Therefore, a collaborative optimization mechanism for noise suppression and geometric correction is constructed to achieve dynamic interaction and collaborative optimization between the two modules:

1. Initial noise suppression: perform initial noise suppression on multimodal fusion features to improve feature quality;

2. Preliminary geometric correction: based on the features after initial noise suppression, perform geometric deformation detection and preliminary correction;

3. Secondary noise suppression: Based on the initially corrected features, perform noise suppression again to eliminate the noise that may be introduced during the geometric correction process;

4. Precise geometric correction: Based on the features suppressed by secondary noise, optimize the geometric correction parameters to achieve precise correction.

Through multiple rounds of collaborative optimization, the maximum effect of noise suppression and geometric correction is ensured, significantly improving the overall accuracy of 3D reconstruction.

## VI. EXPERIMENTAL VERIFICATION AND RESULT ANALYSIS

Experimental Environment and Data Preparation

### A. EXPERIMENTAL ENVIRONMENT

Experimental hardware environment: The CPU is Intel Core i9-13900K, the GPU is NVIDIA RTX 4090 (24GB video

memory), the memory is 64GB DDR5, and the hard drive is 2TB SSD;

Experimental software environment: The operating system is Ubuntu 22.04, the deep learning framework is PyTorch 2.0, the programming language is Python 3.9, and the data processing software includes CloudCompare, ArcGIS, and OpenCV.

### B. EXPERIMENTAL DATA

The experimental data includes publicly available datasets and three types of typical complex mature space actual scene data:

Public datasets: ISPRS Benchmark dataset (including urban and rural scenes), Aerial Image Dataset dataset (including mountain and plain scenes);

Actual scenario data:

1. Mountainous and hilly area: located in the western mountainous area of a certain province, with an area of 5 $km^2$, large terrain undulations, altitude of 500–1500 m, high vegetation coverage, data including 0.5 m resolution optical images, LiDAR point clouds (density 10 points/$m^2$), SAR data (resolution 3 m), and 20 ground control points;

2. Urban dense built-up area: located in the core area of a large city, covering an area of 3 $km^2$, with numerous high-rise buildings and high building density, severe occlusion, and data including 0.3 m resolution optical images, LiDAR point clouds (density of 20 points/$m^2$), SAR data (resolution of 2 m), and 25 ground control points;

3. Wetland Ecological Reserve: Located in a wetland nature reserve with an area of 4 $km^2$, the water and vegetation are interlaced and the lighting conditions are varied. The data includes 0.4 m resolution optical images, LiDAR point clouds (density 8 points/$m^2$), SAR data (resolution 3 m), and 18 ground control points. All data has undergone preprocessing (format conversion, geometric registration, radiometric correction, denoising) to ensure that the data quality meets the experimental requirements.

### C. COMPARATIVE EXPERIMENTAL DESIGN

Select traditional methods and basic deep learning methods as comparison methods to verify the effectiveness and superiority of the proposed method:

Traditional methods: classical SfM MVS method (COLMAP), LiDAR point cloud standalone reconstruction method, optical image and LiDAR point cloud simple fusion reconstruction method;

Basic deep learning methods: MVSNet, Basic U-Net, Transformer model.

Comparative experiments were conducted on public datasets and three types of real-world scenarios, using the same input data and evaluation metrics for all methods to ensure the objectivity of the comparison results.

### D. EXPERIMENTAL RESULTS AND ANALYSIS

#### 1) Public Dataset Experimental Results

The experimental results of the public dataset are shown in Table 4. Compared with traditional methods, the proposed method reduces RMSE by an average of 42.6%, improves SSIM by an average of 38.7%, and increases ORI by an average of 51.3%; Compared with basic deep learning methods, the proposed method reduces RMSE by an average of 28.9%, improves SSIM by an average of 22.5%, and improves ORI by an average of 35.7%. The experimental results show that the proposed method performs well on public datasets, significantly improving the overall accuracy, detail restoration, and occlusion area processing capability of 3D reconstruction.

#### 2) Experimental Results in Actual Scenarios

#### 3) Experimental Results In Mountainous and Hilly Areas

The experimental results in mountainous and hilly areas are shown in Table 2. The RMSE of the proposed method is 0.48m, which is an average decrease of 45.3% compared to traditional methods and an average decrease of 30.2% compared to basic deep learning methods; SSIM is 0.93, which is an average improvement of 40.9% compared to traditional methods and an average improvement of 23.7% compared to basic deep learning methods; The ORI is 87.6%, which is an average improvement of 53.2% compared to traditional methods and an average improvement of 37.8% compared to basic deep learning methods. The experimental results show that the proposed method can effectively address the problems of large terrain undulations, high vegetation coverage, and severe occlusion in mountainous and hilly areas, significantly improving reconstruction accuracy and detail restoration. The Experimental Result Comparison on Public High Solution Aerial Dataset is shown in Table 2:

**TABLE 2.** Experimental Result Comparison on Public High-Resolution Aerial Dataset

| Method | RMSE (m) | MAE (m) | SSIM | ORI (%) |
|---|---|---|---|---|
| COLMAP (Traditional SfM-MVS) | 1.82 | 1.28 | 0.71 | 42.6 |
| LiDAR-only Reconstruction | 1.56 | 1.05 | 0.75 | 58.3 |
| Optics + LiDAR Simple Fusion | 1.35 | 0.92 | 0.79 | 65.8 |
| MVSNet | 1.08 | 0.76 | 0.83 | 72.4 |
| Basic U-Net | 1.15 | 0.81 | 0.81 | 69.7 |
| Transformer | 1.02 | 0.72 | 0.85 | 75.2 |
| Proposed Method | 0.52 | 0.36 | 0.92 | 89.5 |

#### 4) Experimental Results of Densely Populated Urban Areas

The experimental results of densely populated urban areas are shown in Table 3. The RMSE of the proposed method is 0.42m, which is an average reduction of 43.7% compared to traditional methods and an average reduction of 29.5% compared to basic deep learning methods; SSIM is 0.94, which is an average improvement of 39.7% compared to traditional methods and 22.8% compared to basic deep learning methods; The ORI is 91.3%, which is an average improvement of 50.8% compared to traditional methods and

an average improvement of 34.6% compared to basic deep learning methods. The experimental results show that the proposed method can effectively deal with problems such as high-rise building obstruction, scattered buildings, and complex textures in densely populated urban areas, and improve the reconstruction integrity and geometric accuracy of obstructed areas. Experimental Result Comparison in Mountainous and Hilly Areas and Experimental Result Comparison in Dense Urban Built up Areas are shown in Tables 3 and 4:

**TABLE 3.** Experimental Result Comparison in Mountainous and Hilly Areas

| Method | RMSE (m) | SSIM | ORI (%) | GDE (%) |
|---|---|---|---|---|
| COLMAP (Traditional SfM-MVS) | 1.92 | 0.66 | 43.5 | 8.6 |
| LiDAR-only Reconstruction | 1.68 | 0.70 | 59.2 | 6.3 |
| Optics + LiDAR Simple Fusion | 1.45 | 0.74 | 67.8 | 4.8 |
| MVSNet | 1.05 | 0.82 | 70.5 | 3.2 |
| Basic U-Net | 1.12 | 0.79 | 68.3 | 3.7 |
| Transformer | 0.98 | 0.84 | 73.2 | 2.9 |
| Proposed Method | 0.48 | 0.93 | 87.6 | 1.2 |

**TABLE 4.** Experimental Result Comparison in Dense Urban Built-up Areas

| Method | RMSE (m) | SSIM | ORI (%) | SCE (%) |
|---|---|---|---|---|
| COLMAP (Traditional SfM-MVS) | 1.78 | 0.67 | 45.2 | 7.8 |
| LiDAR-only Reconstruction | 1.52 | 0.71 | 61.3 | 5.5 |
| Optics + LiDAR Simple Fusion | 1.38 | 0.75 | 69.4 | 4.2 |
| MVSNet | 0.98 | 0.83 | 75.6 | 2.8 |
| Basic U-Net | 1.05 | 0.80 | 73.1 | 3.3 |
| Transformer | 0.92 | 0.85 | 78.4 | 2.5 |
| Proposed Method | 0.42 | 0.94 | 91.3 | 1.0 |

### 5) Summary of Results Analysis

Based on the comprehensive analysis of publicly available datasets and actual experimental results, the proposed method has the following advantages:

1. The multimodal data deep fusion framework fully explores the complementary features of different data sources, improving the integrity and noise resistance of feature representation;

2. The improved Transformer U-Net network enhances the feature extraction ability of ground edges, fine structures, and occluded areas, improving the accuracy of depth prediction;

3. The collaborative model of adaptive noise suppression and geometric correction effectively filters out noise interference, corrects geometric deformation, and significantly improves reconstruction accuracy;

The proposed method demonstrates excellent performance and robustness across diverse environments ranging from mountainous terrain and dense urban centers to textile industrial clusters, offering broad application prospects for the integrated spatial management of textile manufacturing zones and wetland ecological protection areas.

## VII. OUTLOOK

This paper proposes a deep learning-based 3D remote sensing reconstruction method for complex mature spaces. A unified multimodal feature embedding and attention fusion strategy is designed to fuse LiDAR, SAR, optical and GIS data. The improved Transformer U-Net enhances edge and detail expression. Adaptive noise suppression and geometric correction are collaboratively optimized. Experiments show that the proposed method reduces RMSE significantly and improves detail restoration and robustness. It provides effective technical support for digital twin modeling of textile industrial parks.

This study still has certain limitations:

Model training requires a large amount of annotated data, and obtaining high-quality annotated data in complex mature spaces is costly and time-consuming; The adaptability to extremely complex scenarios such as heavy rainfall, severe occlusion, and earthquake disaster areas still needs further verification; The deployment and application threshold of the model is high, requiring high-performance hardware support, which is not conducive to large-scale promotion.

Future prospects:

Future research can be conducted in the following directions:

1. Optimize the multimodal data fusion framework, adopt lightweight network and parallel computing technology to improve the real-time and efficiency of fusion, and meet the real-time reconstruction requirements;

2. Explore semi supervised learning, weakly supervised learning, and unsupervised learning methods to reduce the model's dependence on annotated data and lower data acquisition costs;

3. Optimize the model structure for extremely complex scenarios (heavy rainfall, severe occlusion, disaster areas), enhance the model's environmental adaptability and anti-interference ability;

4. By developing lightweight models and mobile applications, this approach lowers the deployment threshold to facilitate the large-scale promotion of digital twin technology across textile industrial parks and automated textile production facilities. This ensures that high-precision spatial reconstruction becomes an accessible tool for the real-time management and digital transformation of the textile manufacturing sector.

## AUTHOR CONTRIBUTIONS

Shiman Wang designed, collected and analyzed the data, and drafted the manuscript. Shiman Wang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Shiman Wang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

## CONFLICTS OF INTEREST

The author declares no conflict of interest.

## FUNDING

This research received no external funding.

## ACKNOWLEDGEMENTS

Not applicable.

## REFERENCES

[1] C. Tao, C. Zhang, H. Ji, and J. Qiu, "Fatigue damage characterization for composite laminates using deep learning and laser ultrasonic," Composites Part B: Engineering, vol. 216, Art. no. 108816, 2021, doi: 10.1016/j.compositesb.2021.108816.

[2] S. Wang, T. Xiao, Q. Liu, and H. Zheng, "Deep learning for fast MR imaging: A review for learning reconstruction from incomplete k-space data," Biomedical Signal Processing and Control, vol. 68, Art. no. 102579, 2021, doi: 10.1016/J.BSPC.2021.102579.

[3] W. M. Salama and M. H. Aly, "Deep learning in mammography images segmentation and classification: Automated CNN approach," Alexandria Engineering Journal, vol. 60, no. 5, pp. 4701-4709, 2021, doi: 10.1016/J.AEJ.2021.03.048.

[4] J. Chen, M. Zhou, H. Huang, D. Zhang, and Z. Peng, "Automated extraction and evaluation of fracture trace maps from rock tunnel face images via deep learning," International Journal of Rock Mechanics and Mining Sciences, vol. 142, Art. no. 104745, 2021, doi: 10.1016/J.IJRMMS.2021.104745.

[5] N. Nikzad-Khasmakhi, M. Balafar, M. R. Feizi-Derakhshi, and C. Motamed, "ExEm: Expert embedding using dominating set theory with deep learning approaches," Expert Systems With Applications, vol. 177, Art. no. 114913, 2021, doi: 10.1016/J.ESWA.2021.114913.

[6] D. Yun, A. Abbas, J. Jeon, M. Ligaray, S. S. Baek, and K. H. Cho, "Developing a deep learning model for the simulation of micro-pollutants in a watershed," Journal of Cleaner Production, vol. 300, Art. no. 126858, 2021, doi: 10.1016/J.JCLEPRO.2021.126858.

[7] H. Zhang, Q. Zheng, B. Dong, and B. Feng, "A financial ticket image intelligent recognition system based on deep learning," Knowledge-Based Systems, vol. 222, Art. no. 106955, 2021, doi: 10.1016/J.KNOSYS.2021.106955.

[8] M. Al-Gabalawy, N. S. Hosny, and A. R. Adly, "Probabilistic forecasting for energy time series considering uncertainties based on deep learning algorithms," Electric Power Systems Research, vol. 196, Art. no. 107216, 2021, doi: 10.1016/J.EPSR.2021.107216.

[9] Y. F. Shu, B. Li, X. Li, C. Xiong, S. Cao, and X. Y. Wen, "Deep learning-based fast recognition of commutator surface defects," Measurement, vol. 178, Art. no. 109374, 2021, doi: 10.1016/J.MEASUREMENT.2021.109324.

[10] S. Belagoune, N. Bali, A. Bakdi, B. Baadji, and K. Atif, "Deep learning through LSTM classification and regression for transmission line fault detection, diagnosis and location in large-scale multi-machine power systems," Measurement, vol. 177, Art. no. 109330, 2021, doi: 10.1016/J.MEASUREMENT.2021.109330.

[11] X. A. Wang, J. Tang, and M. Whitty, "DeepPhenology: Estimation of apple flower phenology distributions based on deep learning," Computers and Electronics in Agriculture, vol. 185, Art. no. 106123, 2021, doi: 10.1016/j.compag.2021.106123.

[12] L. Hua, Y. Zhuang, F. Gu, L. Qi, and J. Yang, "AdVLP: unsupervised visible light positioning by adversarial deep learning," Measurement Science and Technology, vol. 32, no. 6, Art. no. 064003, 2021, doi: 10.1088/1361-6501/ABD2DE.

[13] D. Li, S. Peng, Y. Guo, Y. Lu, and X. Cui, "CO2 storage monitoring based on time-lapse seismic data via deep learning," International Journal of Greenhouse Gas Control, vol. 108, Art. no. 103336, 2021, doi: 10.1016/J.IJGGC.2021.103336.

[14] S. Das, K. Kharbanda, R. Raman, and E. Dhas, "Deep learning architecture based on segmented fundus image features for classification of diabetic retinopathy," Biomedical Signal Processing and Control, vol. 68, Art. no. 102600, 2021, doi: 10.1016/J.BSPC.2021.102600.

[15] M. M. Agüero-Torales, J. I. A. Salas, and A. G. López-Herrera, "Deep learning and multilingual sentiment analysis on social media data: An overview," Applied Soft Computing Journal, vol. 107, Art. no. 107373, 2021, doi: 10.1016/J.ASOC.2021.107373.

[16] L. Ilias and I. Roussaki, "Detecting malicious activity in Twitter using deep learning techniques," Applied Soft Computing Journal, vol. 107, Art. no. 107360, 2021, doi: 10.1016/J.ASOC.2021.107360.

[17] P. Huang, Z. Li, C. Wen, J. Lessan, F. Corman, and L. Fu, "Modeling train timetables as images: A cost-sensitive deep learning framework for delay propagation pattern recognition," Expert Systems With Applications, vol. 177, Art. no. 114996, 2021, doi: 10.1016/J.ESWA.2021.114996.

[18] S. Jia, S. Jiang, Z. Lin, N. Li, M. Xu, and S. Yu, "A survey: Deep learning for hyperspectral image classification with few labeled samples," Neurocomputing, vol. 448, pp. 179-204, 2021, doi: 10.1016/J.NEUCOM.2021.03.035.

[19] Z. Li, L. Mihaylova, and L. Yang, "A deep learning framework for autonomous flame detection," Neurocomputing, vol. 448, pp. 205-216, 2021, doi: 10.1016/J.NEUCOM.2021.03.019.

[20] Y. Li, Y. Wang, Y. Zhang, and J. Zhang, "Diagnosis of Inter-turn Short Circuit of Permanent Magnet Synchronous Motor Based on Deep learning and Small Fault Samples," Neurocomputing, vol. 442, pp. 348-358, 2021, doi: 10.1016/J.NEUCOM.2020.04.160.