ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
Received: 8 September 2024; Accepted: 9 January 2025

https://doi.org/10.17559/TV-20240908001975
Original scientific paper

Enhanced Structure-from-Motion 3D Reconstruction through Deep Learning Feature Fusion and Optimization

Dong LI, Gongyun FU, Shunjie YANG*

Abstract: This paper presents Mixum, a novel 3D reconstruction framework for Structure-from-Motion (SfM), which combines traditional feature extraction and matching techniques with deep learning-based optimization. The Mixum framework enhances the accuracy of feature matching and eliminates redundant feature points. Additionally, the integration with PixSfM, a deep-learning accuracy optimization algorithm, further reduces reprojection error and enhances multi-view consistency. Experiments on multiple public datasets reveal that Mixum significantly improves 3D reconstruction density and reduces reprojection error by up to 23%, demonstrating its applicability for complex scenes in applications like cultural heritage preservation, virtual reality, and autonomous navigation.

Keywords: 3D reconstruction; attention mechanism; convolutional neural network; deep learning; feature extraction and matching

1 INTRODUCTION

3D reconstruction techniques are methods of transforming a real-world scene or object into a digital model that a computer can understand. This technology plays an important role in many fields such as virtual reality [1], augmented reality [2], civil engineering [3, 4], autonomous driving [5], cultural heritage protection [6] and medical imaging [7]. It is a key technology for the integration of the real and digital world.

With the development of technology, a variety of methods have been developed for 3D reconstruction. The method [8] that relies on actively emitting light or other radiation to obtain information is called active vision method, and the method that passively receives such as ambient light is called passive vision method [9]. Active vision methods include laser scanning method [10, 11], radar technology [12, 13], kinect technology [14-17], etc., which can directly obtain the depth information of the scene. The passive vision method is more economical. It only needs to use two-dimensional images captured by mobile phones or existing images to reconstruct, and does not require additional equipment, so it is favoured by many solution manufacturers.

Among them, the passive vision method can be further divided into two types of reconstruction methods based on single view [18, 19] and multi-view [20, 21] Single-view reconstruction methods use a single image or video frame for reconstruction, and usually rely on prior knowledge or deep learning techniques to estimate the depth and shape of the object. They are suitable for situations [22] where specific objects exist in the scene or prior information is available. Multi-view reconstruction utilizes the geometric relationship between images or video frames from multiple views to reconstruct. These methods include feature extraction, matching, camera pose estimation and point cloud reconstruction. By matching feature points from different views, the geometric relationship is used to calculate the camera pose and the 3D coordinates of the scene points, so as to realize the reconstruction of 3D scene.

In this paper, we focus on the Structure from Motion (SfM) method for multi-view reconstruction. SfM reconstruction involves feature extraction, matching, image registration, triangulation, and beam adjustment, which is especially suitable for processing a large amount of image data. This method can recover the 3D structure of the scene and the camera position from images taken from multiple angles without relying on expensive hardware, so it has important research value.

In terms of reconstruction accuracy and density, 3D reconstruction techniques can be divided into sparse reconstruction and dense reconstruction [23], and the SfM reconstruction discussed in this paper belongs to the sparse reconstruction category. Sparse reconstruction focuses on the extraction of key feature points to construct a 3D framework, which has the advantages of low computational cost and fast processing speed, so it is very suitable for large-scale scenes, and is often used as a preliminary work for dense reconstruction. However, this method cannot restore the details and textures of the object, and can only roughly outline the shape of the object.

Dense reconstruction can reproduce the fine surface and texture of the object and produce highly realistic models by extracting a large number of pixels and locating them in the 3D space. However, this method is computationally intensive, slow, and requires more resources. Currently, commonly used dense reconstruction techniques include multi-view stereo matching (MVS) [24] and Neural Radiance Fields (Nerf) [25].

Although the traditional SfM method has many advantages, it still has many optimization requirements in complex environments, such as the low accuracy and efficiency of feature extraction and matching, and the low accuracy of pose estimation. However, in recent years, with the in-depth research [26] of deep learning in the field of computer vision, new ideas are provided to solve these optimization requirements. The research on the application of deep learning in 3D reconstruction of SfM is increasing. These research conclusions find that the 3D reconstruction method of SfM based on deep learning can effectively improve the reconstruction effect, especially for complex scenes and environments. Deep learning can automatically learn effective feature representations in data, and can obtain deeper features compared with traditional methods, thereby improving the accuracy and robustness of feature extraction and matching, and can deal with 3D reconstruction tasks in various complex scenes. At the

same time, deep learning can also directly predict the depth or shape information, which simplifies the traditional optimization process. The combination of deep learning and deep learning has brought new research directions to the field of 3D reconstruction, providing more expressive and adaptive 3D scene restoration methods.

At the same time, there are also many challenges in the practical application of 3D reconstruction, such as how to obtain the data set of the reconstruction target, especially when the object to be reconstructed is damaged or destroyed. For this challenge, the Internet datasets provide the possibility to solve it. With the rapid development of mobile Internet and social media, a large amount of historical data has been accumulated on the Internet, especially the photo collections of well-known places or objects. However, factors such as illumination, weather and seasonal changes in online datasets also bring greater challenges to the task of 3D reconstruction. If the network data set can be effectively used for 3D reconstruction, and the diversity and uncertainty of data can be overcome, it will provide great practical value for the practical application of 3D reconstruction.

Based on the traditional 3D reconstruction method of SfM, this paper will study the SfM method based on deep learning optimization, explore the use of deep learning methods to solve the diversity and uncertainty of network data sets, build a 3D reconstruction system of SfM with higher accuracy and robustness, and provide more reliable and efficient solutions for practical application scenarios.

2 LITERATURE REVIEW

In recent years, with the rapid development of deep learning technology, more and more scholars at home and abroad begin to apply deep learning to 3D reconstruction. Deep learning can learn deeper feature representations by training on a large amount of data. With the help of deep learning technology, 3D reconstruction technology has made great progress. At present, deep learning 3D reconstruction methods are mainly divided into two categories, one is the fusion method of deep learning and traditional 3D reconstruction, and the other is the end-to-end deep learning 3D reconstruction method. In the following, the research status of these two categories is introduced in detail.

2.1 Fusion of Traditional 3D Reconstruction with Deep Learning

The method of combining deep learning with traditional 3D reconstruction is a method that integrates the advantages of both. It can not only use the powerful expressive ability and adaptability of deep learning, but also draw on the mature theory and technology of traditional 3D reconstruction.

This paper mainly focuses on the fusion method, and the main idea is that deep learning can be used to optimize and improve a certain link of 3D reconstruction. For example, the deep learning feature extraction methods are mainly D2-Net [27], R2D2 [28], SuperPoint [29], and the deep learning feature matching method is SuperGlue [30]. Chapter 2 will introduce these methods in detail.

In addition, there are also feature matching methods designed based on the characteristics of deep learning, such as the Detector-Free method, which does not rely on traditional feature point detectors to find local features. For example, LoFTR [31] proposes a novel local image feature matching method. This method directly extracts dense local features on two images, and uses the self-attention and cross-attention mechanism in the Transformer architecture to capture feature descriptors with global receptive field and location dependence. LoFTR first establishes pixel-by-pixel dense matching relationships at a coarser scale, and then fine-tunes those high-confidence matches at a finer scale to improve the matching accuracy. Such a method can effectively deal with various challenges in image matching, such as viewpoint variation and illumination variation.

2.2 End-to-End Deep Learning 3D Reconstruction

End-to-end deep learning 3D reconstruction is a method that uses deep neural networks to directly generate 3D models from input images or videos. It does not require manually designed features or intermediate representations, but realizes the task of 3D reconstruction through end-to-end learning of the network. The advantages of this method are that it can make full use of the ability of deep learning, reduce human intervention and assumptions, directly omit the intermediate steps of traditional 3D reconstruction to reduce errors, and improve the efficiency and quality of 3D reconstruction. The challenge of this method is how to design an appropriate network structure and loss function, and how to deal with large-scale and complex 3D data. Currently, end-to-end deep learning 3D reconstruction has made some progress.

The DeMoN [32] network was proposed at CVPR 2017, and it jointly estimates depth and camera motion from two consecutive images by training a convolutional network. DeMoN is mainly composed of three parts, which are bootstrap network, iterative network and refinement network. The Bootstrap network takes an image pair as input and outputs the initial depth and motion estimates, the iterative network improves the existing depth, surface normal, and motion estimates, and the refined network upsamples the prediction results to the original input resolution.

BA-NET [33] addresses the problem of SfM through a novel approach called feature-metric Bundle Adjustment, which explicitly enforces multi-view geometric constraints in the form of Feature Metric errors. The whole process is differentiable, so the network can learn suitable features, which makes the Bundle Adjustment (BA) problem easier to solve. The essence of the BA problem is to optimize the camera parameters and 3D point positions by minimizing the reprojection error between images. It is a nonlinear least squares problem. In addition, the proposed method introduces a novel depth parameterization method to recover the dense depth per pixel. Its network first generates several base depth maps from the input image, and optimizes the final depth to a linear combination of these base depth maps by the feature metric BA. The basic depth map generator is also learned through end-to-end training, and the whole system combines domain knowledge and deep learning well.

DeepSFM [34] learns from the idea of traditional beam adjustment, and constructs a cost volume through two deep learning models for depth map and pose estimation, respectively. In DeepSFM, 2D CNN is used to extract photometric features to construct the cost volume. The initial source depth map and camera pose are used to introduce photometric and geometric consistency. While a series of 3D CNN layers are applied to D-CV and P-CV, respectively, and then a context network and depth regression operation are applied to generate the predicted depth map of the target image.

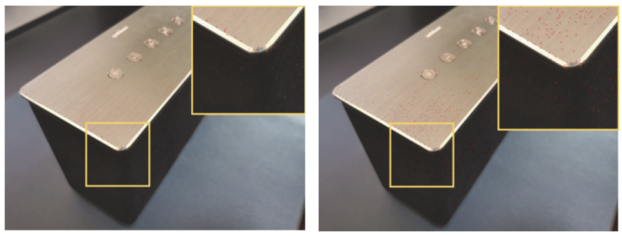
2.3 Summary

From the above research status at home and abroad, it can be seen that deep learning has shown remarkable achievements in the field of 3D reconstruction, but there are still some shortcomings at this stage. For example, the stability of current deep learning methods is limited by the quality of the training data set, which is easily affected by uncertainty and noise, resulting in lack of generality and stability. In contrast, traditional methods have shown good stability on general datasets after years of development, but still have limitations on complex datasets. Therefore, based on the research status at home and abroad, this paper will combine the advantages of deep learning, integrate traditional methods, and focus on the SfM optimization algorithm based on deep learning.

3 PROBLEM DESCRIPTION

3.1 Feature Point Extraction

Traditional algorithms are mainly based on local information to obtain feature descriptors, and have limited expression ability for high-level abstract semantic features. Therefore, they cannot perform well on some complex scene datasets, such as weak textures, repeated textures, smooth surfaces or blurred cases, and their feature description ability is relatively shallow, so they cannot effectively deal with occlusion, season and illumination changes. However, the deep learning algorithm can extract deeper and more comprehensive feature information through multi-layer convolution, which can effectively deal with the shortcomings of traditional algorithms.



a) SIFT feature extraction results b) SuperPoint feature extraction results

Figure 1 Comparison of feature extraction results

As shown in Fig. 1, when the target is an object with repetitive textures or smooth surfaces, SuperPoint outperforms SIFT in feature extraction.

Therefore, a multi-feature fusion SfM 3D reconstruction framework supporting deep learning and traditional feature extraction and matching algorithms can be designed to optimize the effect of feature extraction and matching, and obtain the SfM 3D reconstruction point cloud results and related evaluation index data. Finally, by selecting datasets with seasonal, illumination, climate and other characteristics changes, comparative experiments can be carried out to verify whether the framework can solve the above problems and optimize the results of SfM 3D reconstruction.

3.2 Accuracy Issues

Traditional image matching methods usually detect feature points once in each image, which easily leads to inaccurate positioning of feature points and poor repeatability, which greatly affects the accuracy of the final 3D reconstruction results. At the same time, the traditional SfM method is mainly based on the geometric alignment of feature points to estimate the camera pose and 3D space points, which is difficult to ensure the consistency of camera pose and scene geometry, especially in the presence of a lot of detection noise and appearance changes. In addition, traditional SfM 3D reconstruction methods usually use hand-designed feature points and descriptors for matching and reconstruction operations, which may significantly reduce the representation and generalization ability of features under complex data sets.

4 SOLUTION

Previously, traditional feature extraction algorithms were widely used due to their robustness and reliable feature representation ability. However, in the scene with no significant local features such as weak texture, texture repetition and image blur, the effect of traditional algorithms is not good, so the 3D reconstruction results of SfM are also affected.

In order to solve the above problems, we first propose a multi-feature fusion SfM framework called Mixum. In this framework, deep learning and traditional feature data are fused, and redundant low reliable repeated points or nearly repeated points are eliminated. By integrating the advantages of different feature extraction and matching algorithms and complementing each other, more comprehensive and richer feature information can be obtained, so as to improve the effect of feature point extraction and matching, and then improve the accuracy and robustness of 3D reconstruction.

Secondly, this paper also combines the Mixum framework with the deep learning accuracy optimization algorithm PixSfM. This algorithm is optimized for the error in the 3D reconstruction process, and can further improve the accuracy of the reconstruction results of the Mixum framework when combined with the Mixum framework.

4.1 Multi-Feature Fusion SfM Framework Mixum

This paper proposes a 3D reconstruction framework of SfM based on multi-feature fusion. The framework selects two different feature extraction and matching methods to extract and match the features of the dataset images respectively, and then fuses the obtained feature extraction data and feature matching data; at this time, there will be redundant points in direct fusion data. Therefore, it is necessary to eliminate the low reliable duplicate and

approximately duplicate feature point data, and then send it into the subsequent incremental SfM 3D reconstruction process for subsequent image registration, triangulation, BA optimization and other steps. Finally, the point cloud model of SfM 3D reconstruction and the evaluation index data of the result are obtained.

The diagram of multi-feature fusion Mixum framework is shown in Fig. 2 below.

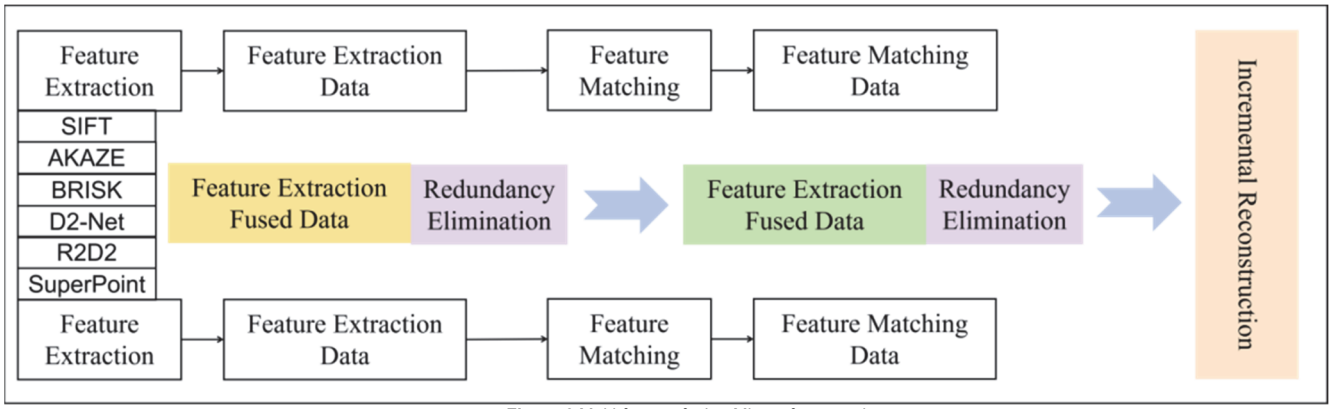


Figure 2 Multi-feature fusion Mixum framework

Next, this section will elaborate the specific implementation of the solution from two aspects: feature extraction data fusion and feature matching data fusion.

(1) Feature extraction data fusion

At this stage, the primary focus is on effectively fusing the feature extraction data from two different methods, which requires a screening and preprocessing step prior to the actual data fusion. The feature extraction data primarily includes the locations of feature points and their corresponding descriptor information. The screening preprocessing involves several key steps. First, it identifies and isolates duplicate feature points and approximately duplicate feature points. Following this, it classifies these points into their respective categories. Finally, using the existing feature matching information, it eliminates low-reliability points to ensure that the subsequent data fusion is based on high-quality and reliable feature points. This process enhances the robustness and accuracy of the fused feature dataset.

Approximately duplicate feature points refer to those extracted from the same image by two different feature extraction methods, which are very close to each other but not necessarily located at the exact same pixel position. In practice, such points can be considered duplicate feature points and eliminated during processing. As illustrated in Fig. 3, this scheme identifies nearly duplicate feature points based on a position approximation threshold. The circle in the figure represents the threshold range. If two points are within this threshold range and share the same position, they are considered repeated feature points. If the positions are different but still fall within the threshold range, they are classified as approximately repeated feature points. This approach ensures that redundant feature points are effectively managed during the extraction and matching process.

The expression of approximately repeated feature points is shown in the Eq. (1), where $p_1 = (x_1, y_1)$ represents the position of feature points in method 1, $p_2 = (x_2, y_2)$ represents the position of feature points in method 2, and R is the radius of the approximate domain. When the distance between two points is within the approximate domain, it means that two points are approximately repeated feature points.

$$(x_2 - x_1)^2 + (y_2 - y_1)^2 \le R^2 \tag{1}$$

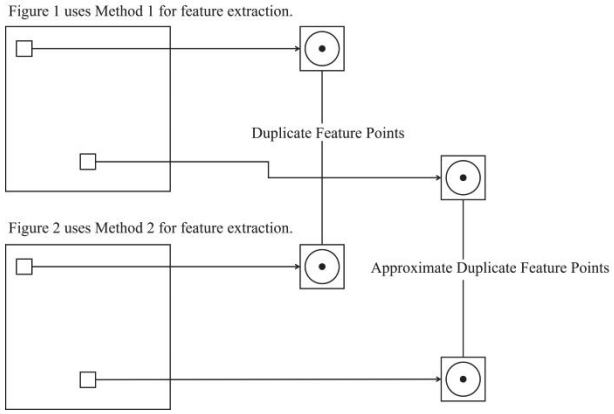


Figure 3 Schematic diagram of repeated and approximately repeated feature points

Next, the matching information of duplicate and nearly duplicate feature points with other images is queried. From these, only the feature points with the best matching results are retained, while the others are classified as low-reliability points and subsequently eliminated. This elimination process also includes removing the corresponding matching data associated with these low-reliability points to ensure the integrity and accuracy of the remaining data. This step is critical for refining the feature set and improving the reliability of the feature extraction and matching process.

The expression of low reliable points is given in the Eq. (2), where p_{low} is low reliable point, $M(p_1)$ is the matching number of feature points p_1, $M(p_2)$ is the matching number of feature points p_2, and min is the feature point with less matching number between the two.

$$p_{\text{low}} = \min(M(p_1), M(p_2)) \tag{2}$$

Fig. 4 presents a schematic diagram illustrating low-reliability points in the feature extraction and matching process. In the diagram, points P_1 and P_2 represent repeated or approximately repeated feature points identified through

feature extraction and matching methods 1 and 2. Solid lines depict the connections between matching points, while dotted lines indicate the connections between repeated or approximately repeated points. The identification of low-reliability points follows the principle that the point with fewer matching lines between the two is categorized as a low-reliability point. It is worth noting, however, that this criterion is not the sole method for determining low-reliability points; other conditions or criteria may be employed depending on the specific requirements of the analysis or application at hand.

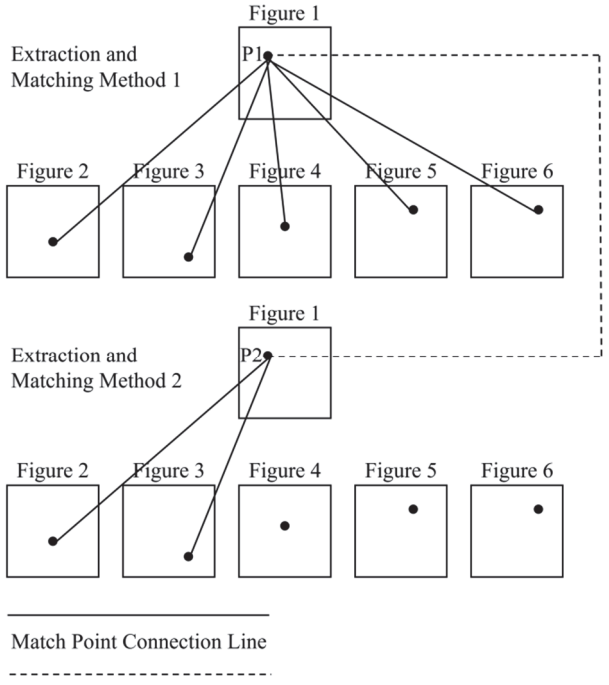


Figure 4 Schematic representation of low reliability points

The following is the flow of the feature extraction data fusion algorithm, shown in Algorithm 1.

Table 1 Feature extraction data fusion algorithm

Algorithm 1 Feature Extraction Data fusion algorithm

Input:

Method 1 Feature extraction data, matching data $P_1 = \{p_1^1, p_1^2, ..., p_1^n\}$ $QM_1 = \{M_1^1, M_1^2, ..., M_1^n\}$.

Method 2 Feature extraction data, matching data $P_2 = \{p_2^1, p_2^2, ..., p_2^m\}$ $QM_2 = \{M_2^1, M_2^2, ..., M_2^n\}$.

$p_x^i = \{x_x^i, y_x^i\}$ is the feature point location, feature matching set $M_x^i = \{p_x^i p_x^{k1}, p_x^i p_x^{k2}, ..., p_x^i p_x^{kn}\}$.

The subscript x is the sequence number of the extraction matching method, and the *superscript i* and k denote the sequence number of the corresponding image set.

Output: Fused feature point data $P_e = \{p_e^1, p_e^2, ..., p_e^t\}$

1: Ensemble $P(P_1, P_2)$
2: for i = 1 to n **do**
3: for j = 1 to m **do**
4: if p_2^j not in P_2 **or** p_1^i not in P_1 **then**
5: Continue
6: if p_1^i **and** p_2^j isapprox point **then**
7: if length (M_1^i) < length (M_2^j) **then**
8: **remove** p_1^i **from** P_1
9: **remove** M_1^i **from** QM_1
10: **else**
11: **remove** p_2^j **from** P_2
12: **remove** M_2^i **from** QM_2
13: $P_e = P_1$ **concat** P_2
14: return P_e

(2) Feature matching data fusion

Since the matching data with low reliability points have been eliminated in the data fusion stage of feature extraction, the feature matching data are mainly directly fused in this stage.

Table 2 Feature matching data fusion algorithm

Algorithm 2 Feature Extraction Data fusion algorithm

Input:

Method 1 march the data $QM_1 = \{M_1^1, M_1^2, ..., M_1^n\}$

Method 2 march the data $QM_2 = \{M_2^1, M_2^2, ..., M_2^n\}$

Feature matching collection

$M_x^i = \{p_x^i p_x^{k1}, p_x^i p_x^{k2}, ..., p_x^i p_x^{kn}\}$

The subscript x is the sequence number of the extraction matching method, and the *superscript i* and k refer to the sequence number of the corresponding image set.

Output: Fusion feature matching data

$QM_e = \{p_e^1, p_e^2, ..., p_e^t\}$

1: Ensemble $M(QM_1, QM_2)$
2: $QM_e = QM_1$ concat QM_2
3: return QM_e

When the above two steps are completed, the feature extraction and matching data after the fusion and elimination of redundant data need to be passed to the next stage for the subsequent incremental reconstruction step. After registering each image, triangulation and B-beam adjustment are performed, and the sparse reconstruction result is finally obtained.

4.2 Accuracy optimization of Mixum combination PixSfM

To solve the accuracy optimization problem of the algorithm, the deep feature alignment optimization method can be used. First, all the matched 2D feature points can be jointly fine-tuned before SfM to maximize their feature alignment among multiple views, thereby improving the accuracy and repeatability of feature points. At the same time, the 3D points and camera pose can be fine-tuned after SfM to maximize their feature alignment among multiple views, thereby improving the accuracy of 3D reconstruction and visual localization. At the same time, a deep learning network can be used to extract dense deep features from images, which can maintain consistency in different views and scenes, thereby improving the expression and generalization ability of features.

This paper proposes that the Mixum framework can be combined with the deep learning accuracy optimization algorithm to further improve the accuracy of SfM 3D reconstruction results on the basis of multi-feature fusion Mixum framework optimization. The combined scheme architecture is shown in Fig. 5 below.

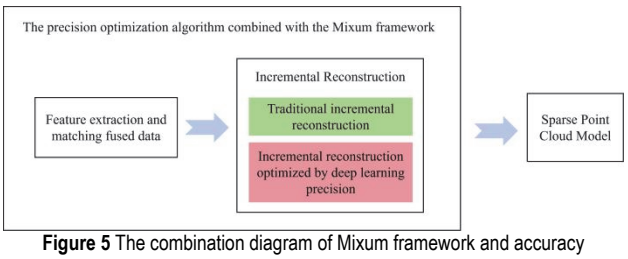


Figure 5 The combination diagram of Mixum framework and accuracy optimization algorithm

PixSfM (Pixel-Perfect-SfM) optimization method was proposed by Lindenberger and Sarlin et al. at ICCV 2021, which optimizes feature points, camera poses, and 3D points by directly aligning deep features.

As shown in Fig. 6, this method is optimized in two main steps. One is to use the depth Feature metric to optimize the position of feature points after feature matching, and the other is to perform BA optimization by the depth feature metric error in the incremental process.

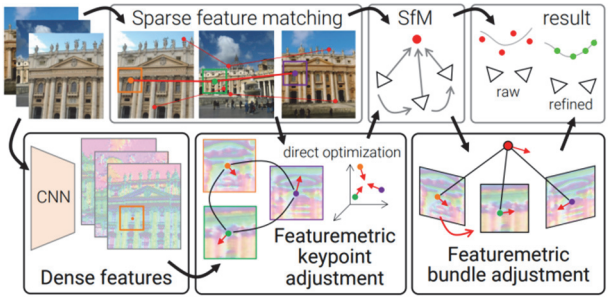


Figure 6 Pixel-Perfect-SfM optimization flowchart

Specific implementation steps are as follows:

(1) A pre-trained CNN is used to extract dense deep features from images that are consistent across viewpoints and scenarios.

(2) A cost function based on feature metric is used to jointly fine-tune all the matched 2D feature points to maximize their feature alignment across multiple views. This step can be performed before SfM to improve the accuracy and repeatability of feature points.

The formula for the Featuremetric is shown in Eq. (3), where, x_i is a point in the ith view, $f_i(x_i)$ is the depth feature of that point, and $\|\cdot\|$ is the Euclidean norm. The smaller the value of the feature metric, the higher the feature alignment between multiple views.

$$F(x_1, x_2, \ldots, x_n) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \| f_i(x_i) - f_j(x_j) \|^2 \quad (3)$$

The objective function for feature point fine-tuning is shown in the Eq. (4), where x_i^0 is the initial feature point position in the ith view and λ is a regularization coefficient. This objective function aims to minimize the feature metric while maintaining the initial keypoint position.

$$\min_{x_1, x_2, \ldots, x_n} F(x_1, x_2, \ldots, x_n) + \lambda \sum_{i=1}^{n} \| x_i - x_i^0 \|^2 \quad (4)$$

(3) Use COLMAP or another SfM system to perform camera pose estimation and 3D point reconstruction based on the fine-tuned keypoints. This step can be done using known camera poses or from scratch.

(4) Using the same feature metric cost function, the 3D points and camera poses are further fine-tuned such that their feature alignment across multiple views is maximized. This step can be performed after SfM to improve the accuracy of 3D reconstruction and visual localization.

Among them, the objective function of 3D point and camera pose fine-tuning is given in Eq. (5) where X is the coordinates of 3D points, R_i and t_i are the camera rotation matrix and translation vector of the ith view, x_i is the feature point position in the ith view, and λ is a regularization coefficient. This objective function aims to minimize the feature metric while preserving the projection error.

$$\min_{X, R_1, t_1, \ldots, R_n, t_n} \sum_{i=1}^{n} F(X, R_i X + t_i, \ldots, R_n X + t_n) + \lambda \sum_{i=1}^{n} \| R_i X + t_i - x_i \|^2 \quad (5)$$

PixSfM can effectively calibrate the errors caused by detection noise or appearance changes, and improve the accuracy of reconstruction.

In this paper, PixSfM is selected as a scheme to optimize the accuracy of 3D reconstruction through deep features, and it is combined with the Mixum framework. In the future, this combination method will be compared to obtain the optimization effect of the combination method on the 3D reconstruction results of SfM.

The combination of Mixum framework and PixSfM algorithm is shown in Fig. 7 below.

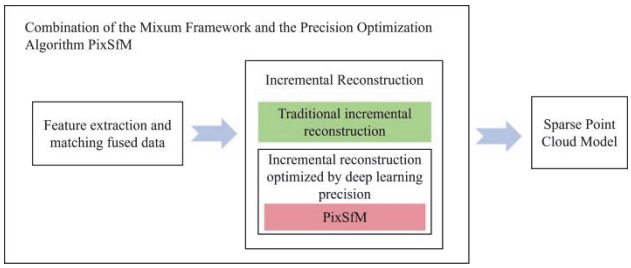


Figure 7 How the Mixum framework and PixSfM combine

5 EXPERIMENTAL RESULTS

In this section, three public datasets are selected to conduct comparative experiments on the above algorithms, so as to obtain the performance data of the corresponding algorithms. This experiment can be divided into three comparison experiments. The first is the comparison experiment of feature extraction and matching algorithms, which aims to obtain the 3D reconstruction effect data of single algorithm on the data set. The second is the comparison experiment of multi-feature fusion optimization, which aims to obtain the optimization benefits of 3D reconstruction effects before and after using the multi-feature fusion Mixum framework, and selecting

different algorithms for fusion. The third is the accuracy optimization comparison experiment, the purpose is to obtain the optimization data of 3D reconstruction effect before and after using the accuracy optimization combination of Mixum framework and PixSfM.

5.1 Datasets

In this paper, three classic data sets are selected for experiments, and the basic information is shown in the table below:

Table 3 Datasets basic information

Data Sets	Number of images	Resolution	Features
South Building	128	3094 × 2312	The same time period, the same equipment, the same lighting, the same season
British Museum	176	All different	Different equipment, different time of day shooting, lighting, seasonal changes
Sacre Coeur	281	All different	Different equipment, different time of day shooting, lighting, seasonal changes

5.2 Evaluation indicators

(1) Number of registered images

The number of registered images indicates the number of images that can successfully locate the camera pose during the reconstruction process and are used for reconstruction. A higher number of registered images means that more data is used for reconstruction, which usually provides a more detailed and complete model.

(2) The number of 3D points

3D points refer to the number of points in the reconstructed 3D point cloud. A higher number of points usually means a denser point cloud is obtained, which indirectly indicates that the scene is more detailed.

(3) Average point trajectory length

Average Track Length (ATL) is a metric that measures the consistency of recognizing and tracking a 3D point from multiple views in 3D reconstruction. Specifically, it represents the average length of the 2D trajectory projected by a point across all views. As shown in the Eq. (6), where N denotes the number of 3D points and L_i is the trajectory length of the ith 3D point, that is, how many different views the 3D point is observed in.

$$ATL = \frac{1}{N}\sum_{i=1}^{N} L_i \tag{6}$$

A longer average point trajectory length means that each 3D point has a corresponding 2D point in more images, which can improve the estimation accuracy of the 3D point position. In contrast, a short average point trajectory length may imply inconsistencies in the matching of feature points in different views, which may be due to inaccurate feature point matching, image noise, viewpoint change restrictions, or other interfering factors. This metric provides indirect information about the quality of feature point matching and the accuracy of reconstructed points.

(4) Average reprojection error

The average reprojection error is the average distance error between the estimated position and the actual point reprojected to each image calculated by the pose for each 3D point. A smaller average reprojection error usually indicates a more accurate and precise reconstruction result. This indicates that the estimated positions of the reprojected points are very close to the corresponding actual feature points, which indirectly indicates that the overall structural accuracy is higher.

For the projection of N 3D points on M images, the average reprojection error E can be expressed as the Eq. (7), where P_i is assumed to be the coordinates of 3D point i, p_{ij} is the 2D projection position of the point on image j, and $\hat{p}_{ij}$ is the predicted projection position obtained from 3D reconstruction and camera parameters.

$$E = \frac{1}{N \cdot M}\sum_{i=1}^{N}\sum_{j=1}^{M}\left\| p_{ij} - \hat{p}_{ij} \right\| \tag{7}$$

(5) Visualization effect

By visualizing the point cloud of the 3D reconstruction result, the density, accuracy, and completeness of the point cloud can be intuitively evaluated. Good SfM reconstruction results should produce accurate and complete point clouds that reflect real-world scenes.

5.3 Comparison Experiment of Feature Extraction and Matching Algorithms

This section will conduct comparative experiments on several feature extraction and matching algorithms mentioned above to obtain the performance data of different algorithms on different public datasets.

(1) Comparative experiments of traditional feature extraction algorithms

This comparison experiment will compare three traditional feature point extraction methods, aiming to find the traditional feature extraction method that is most suitable for subsequent comparison with the feature extraction method of deep learning for 3D reconstruction. Since the main purpose of this study is to find a method that is more robust to illumination and seasonal changes, this experiment mainly selects pictures with large differences in illumination, Angle and scale as the data source of the comparison experiment.

The number of feature point extraction and extraction time are selected as the indirect evaluation indicators, and then the repeatability and scale invariance are selected as the key evaluation indicators.

Repeatability refers to running the same feature extraction algorithm at different times or on different images, and the feature description obtained by the same feature point can maintain a certain degree of consistency and stability. In simple terms, it should produce similar feature descriptions under similar scene conditions.

Scale invariance, on the other hand, refers to the ability to maintain consistent performance and representation when processing objects or scenes at different scales in an image. This means that no matter how large or small the size of the object is in the image, such as how far away it

is when taking a photo, the algorithm will be able to correctly detect, recognize or describe these objects without being affected by scale changes.

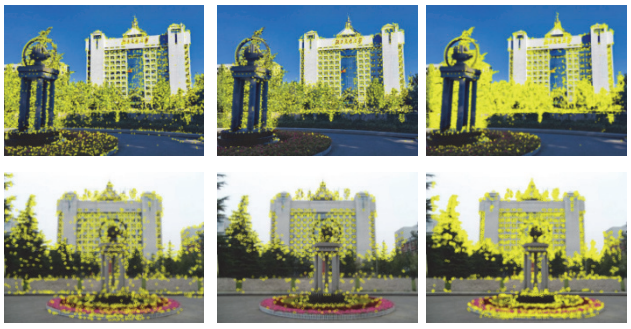
The experimental results of indirect evaluation indicators for traditional feature extraction methods are shown in Tab. 4 below.

Table 4 Comparison table of indirect evaluation indicators for traditional feature extraction methods

Data	Methods	Image Dimensions	Number of feature points	Extraction time consuming
Example1	SIFT	1080 × 809	8339	0.105
	AKAZE	1080 × 809	5369	0.053
	BRISK	1080 × 809	16046	0.151
Example2	SIFT	600 × 450	2245	0.034
	AKAZE	600 × 450	1194	0.021
	BRISK	600 × 450	3851	0.037

The quantitative dimension was extracted from the feature points for comparison, and the experimental results were compared BRISK > SIFT > AKAZE. In the dimension of extraction time, the experimental results AKAZE < SIFT < BRISK were compared. From the research direction of this paper, the weight of extraction effect is higher than that of extraction speed.

In the next step, visual comparison experiments will be carried out to evaluate the repeatability and scale invariance. The results of the visual comparison experiment are shown in Fig. 8 below:



a) SIFT b) AKAZE c) BRISK

Figure 8 Visual comparison of three feature extraction methods

BRISK extracts a lot of feature points, but the distribution is too concentrated and densely distributed, while the better feature extraction results are evenly distributed. At the same time, it can be seen from the figure that compared with SIFT, the flower bed and wall parts fail to extract good feature points.

From the perspective of visual comparison, SIFT is better than BRISK. Compared with BRISK, AKAZE extracted the least number of feature points. Like BRISK, AKAZE failed to extract well-distributed feature points in some areas such as flower beds and walls.

In general, from the perspective of visualization, the feature extraction effect is SIFT > BRISK > AKAZE.

In order to further prove the results of the above visual comparison map analysis, the next step is to introduce the feature matching effect comparison experiment to further determine the repeatability and scale invariance of the feature extractor. The matching experimental results of three traditional feature extraction algorithms are shown in Fig. 9 below.

a) SIFT feature matching results

b) AKAZE feature matching results

c) BRISK feature matching results

Figure 9 Comparison of matching results of three kinds of feature extraction

From the feature matching results, it can be seen that the current test images have large differences in illumination, season, and scale changes. Under the same feature matching method, the number of feature matching is SIFT > BRISK > AKAZE.

It can be seen that SIFT has the best robustness to illumination, season and scale changes in the comparison of traditional methods, so SIFT is selected to compare with deep learning feature extraction algorithms.

(2) Comparison experiment of 3D reconstruction between traditional and deep learning algorithms

This experiment will compare the 3D reconstruction effect of SIFT, D2-Net, R2D2 and SuperPoint. The evaluation indicators include the number of registered images, 3D points, average point trajectory length, and average reprojection error.

Experimental results see table, the table method named used abbreviation, corresponding to nearest neighbor (NN) matching method, SuperPoint (SP), SuperGlue (SG).

The comparison results of 3D reconstruction effect evaluation index data in this experiment are shown in Tab. 5 below.

Contrast can be seen from evaluation index data, South - Building on the special gathering data set, the characteristics of data set for collection at the same time use the same equipment.

Judging from the results, SIFT overall performance is good, has the lowest average projection error, and reconstruction of 3D points out more, and the other three

methods of deep learning feature extraction is the average weight projection error is higher, but keep in mind, the D2 - Net reconstruction of 3D points out the most, the subsequent can combine visual point cloud to do further analysis.

And for the other two Internet data collection of British Museum and Sacre Coeur, the characteristics of data set for the use of different equipment acquisition, and illumination, have significant changes in the weather, season, images with different noise at the same time, such as photo crowd, shade, etc.

Table 5 Comparison table of 3D reconstruction effect evaluation index data

Dataset	Methods	Number of registered image	3D points	Average point track length	Average projection error
South Building	SIFT+NN	128	35519	5.889	0.882
	D2-Net+NN	128	51249	4.095	1.393
	R2D2+NN	128	25299	6.465	1.369
	SP+SG	128	35529	4.929	1.374
British Museum	SIFT+NN	176	8073	11.968	0.897
	D2-Net+NN	160	7774	8.740	1.292
	R2D2+NN	173	10419	22.765	0.923
	SP+SG	176	18396	10.315	1.419
Sacre Coeur	SIFT+NN	280	16686	12.143	0.844
	D2-Net+NN	268	15449	9.719	1.291
	R2D2+NN	281	15725	20.337	0.766
	SP+SG	281	31633	11.935	1.361

From the results, SuperPoint extracts the largest number of 3D points in the two datasets.

Therefore, according to the preliminary judgment of the current implementation data comparison, the SuperPoint deep learning feature extraction algorithm is better than SIFT when illumination, weather and seasonal changes are large. In conventional light, weather, seasonal change on small data sets, SIFT traditional methods still have a better extraction effect.

The following is a comparison from the visual point cloud effect, and we can judge the reconstruction effect by subjective viewing.

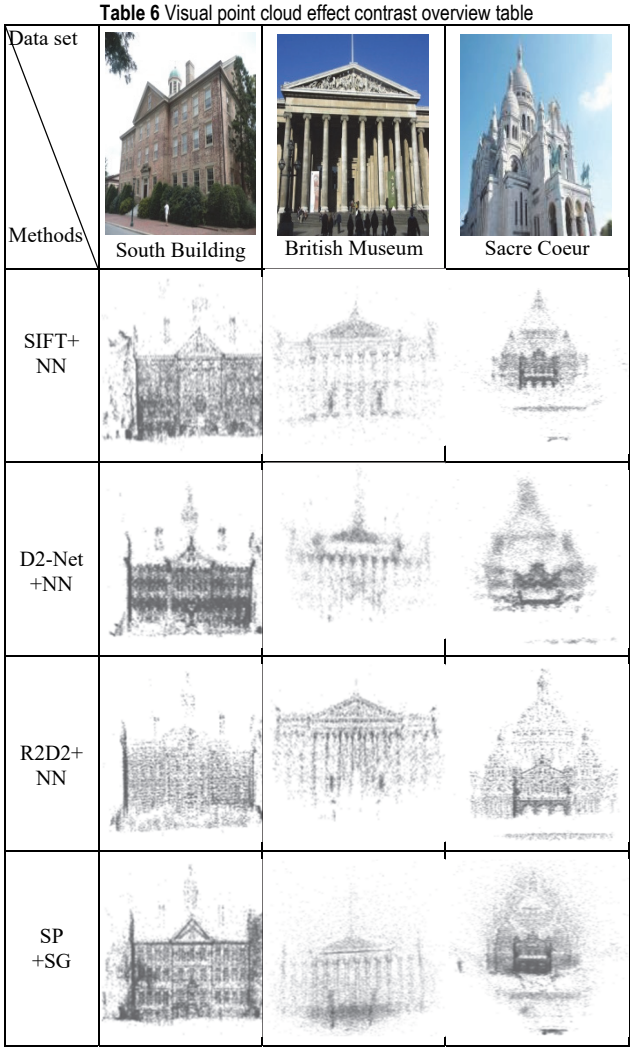
Visual point cloud results are shown in Tab. 6 below.

From the visual effect of point cloud, as you can see, in the South-Building D2-Net, although reconstruction of 3D points out, most are densely concentrated mainly on some repeat position, such as Windows, so most of the three-dimensional point invalid. From the reconstruction effect, SIFT > SuperPoint > D2-Net > R2D2.

Comparing the next data set on the Internet to check the reconstruction effect, SuperPoint > R2D2 > SIFT > D2-Net. SuperPoint has the best effect, and it can be seen that the 3D points of the feature parts of the original object are the most perfect, while D2-Net has the worst effect, the 3D point positions are particularly scattered, and no good point cloud is formed to describe the original object.

From the contrast experiment result shows that the data set South Building, lighting, weather, seasonal factors, such as relatively consistent, the original object representation does not appear as too big change, the effect is better than traditional methods SIFT deep learning algorithm. In data collection of British Museum and the Sacre Coeur, factors such as light, weather, and season change are bigger, the original object representation in the larger changes, the deep learning feature extraction methods SuperPoint reconstruction effect is best.

Table 6 Visual point cloud effect contrast overview table



5.5 Comparative Experiment of Multi-Feature Fusion

In this section before and after the experiment method of using multiple feature fusion Mixum framework comparison experiment, the aim is to get the framework multiple features fusion method to improve data reconstruction effect. In this experiment, D2-Net, R2D2 and SuperPoint will be compared with SIFT fusion.

Data in Tab. 7 for the multiple characteristics of the fusion results of experiment data before and after optimization. Which method the @ said fusion in parentheses, @ left as a benchmark method, the fusion method is on the right and brackets for the use of multiple features fusion Mixum framework after value evaluation index of the experimental data.

From comparative data, as you can see, using multi-model fusion method, 3D reconstruction on the number of registered image, 3D points are greatly improved, and the average weight projection error is reduced by about 12%. After using multi-model fusion optimization method, the generated sparse point cloud point location is more accurate, and can extract the point more.

So from the multi-model integration optimization of contrast experiment, the multi-model fusion method in 3D reconstruction public data sets of South Building, British Museum, Sacre Coeur can enhance the effect of 3D reconstruction, 3D reconstruction after optimization can effectively improve feature points, at the same time let the reconstruction results more accurate.

Table 7 More characteristic model integration optimization of contrast table

Dataset	Methods	Register Image Count	3D points	Average point track length	Average reprojection error
South Building	D2-Net+NN	128 (128).	51249 (88938)	4.095 (4.782)	1.393 (1.196)
	Ours (D2-Net+NN@SIFT+NN)				
	R2D2+NN	128 (128).	25299 (60735).	6.465 (6.136)	1.369 (1.087)
	Ours (R2D2+NN@SIFT+NN)				
	SP+SG	128 (128).	35529 (72741).	4.929 (5.367)	1.374 (1.140)
	Ours (SP+SG@SIFT+NN)				
British Museum	D2-Net+NN	160 (176).	7774 (16779).	8.740 (10.083)	1.292 (1.113)
	Ours (D2-Net+NN@SIFT+NN)				
	R2D2+NN	173 (176).	10419 (18277).	22.765 (18.273)	0.923 (0.920)
	Ours (R2D2+NN@SIFT+NN)				
	SP+SG	176 (176).	18396 (26073).	10.315 (10.947)	1.419 (1.274)
	Ours (SP+SG@SIFT+NN)				
Sacre Coeur	D2-Net+NN	268 (281).	15449 (32682).	9.719 (10.932)	1.291 (1.085)
	Ours (D2-Net+NN@SIFT+NN)				
	R2D2+NN	281 (281).	15725 (32517).	20.337 (16.165)	0.766 (0.816)
	Ours (R2D2+NN@SIFT+NN)				
	SP+SG	281 (281).	31633 (47913).	11.935 (12.103)	1.361 (1.196)
	Ours (SP+SG@SIFT+NN)				

5.6 Comparison Experiment of Precision Optimization

The experiments in this section will compare the deep learning accuracy optimization algorithm PixSfM before and after use, in order to obtain the accuracy optimization effect of the accuracy optimization algorithm on the SfM framework Mixum. As shown in Tab. 8, the method items not in parentheses in the former are the methods using only the Mixum framework, and the latter are the methods using Mixum combined with PixSfM in parentheses. Among them, PixSfM is abbreviated as PPS in the table, #PPS means that this method applies the PixSfM method, @ means fusion, and before and after @ are the specific algorithms corresponding to the corresponding fusion.

By comparing the data before and after optimization, it can be seen that the average reprojection error has a very significant improvement after accuracy optimization, and the average reduction is 23% in general. From preliminary data, the accuracy of the optimization method can effectively improve the accuracy of 3D reconstruction point.

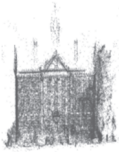
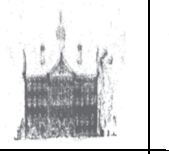
The following is a comparison of the visual point cloud effect. The top and bottom two pictures in the table show the effect before and after using accuracy optimization, and the one with #PPS represents the result of using accuracy optimization method.

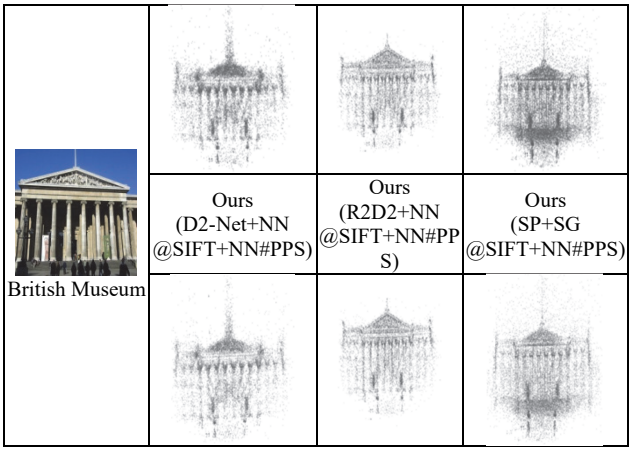
Table 8 Accuracy optimization effect comparison table

Dataset	Methods	Register Image Count	3D points	The average point trajectory length	Average reprojection error
South Building	D2-Net+NN@SIFT+NN	128 (128).	88938 (86003).	4.782 (5.277)	1.196 (0.566)
	Ours (D2-Net+NN@SIFT+NN#PPS)				
	R2D2+NN@SIFT+NN	128 (128).	60735 (59126).	6.136 (6.426)	1.087 (0.617)
	Ours (R2D2+NN@SIFT+NN#PPS)				
	SP+SG@SIFT+NN	128 (128).	72741 (68556).	5.367 (5.837)	1.140 (0.764)
	Ours (SP+SG@SIFT+NN#PPS)				
British Museum	D2-Net+NN@SIFT+NN	176 (176).	16779 (16893).	10.083 (9.967)	1.113 (0.964)
	Ours (D2-Net+NN@SIFT+NN#PPS)				
	R2D2+NN@SIFT+NN	176 (176).	18277 (19190).	18.273 (17.286)	0.920 (0.935)
	Ours (R2D2+NN@SIFT+NN#PPS)				
	SP+SG@SIFT+NN	176 (176).	26073 (26440).	10.947 (10.736)	1.274 (1.227)
	Ours (SP+SG@SIFT+NN#PPS)				

As shown in Tab. 9, from the comparison results, the point cloud position after accuracy optimization is more accurate.

Table 9 Accuracy optimization effect point cloud visualization comparison table

The data set	Methods of comparison		
South Building	D2-Net+NN @SIFT+NN	R2D2+NN @SIFT+NN	SP+SG @SIFT+NN
			
	Ours (D2-Net+NN @SIFT+NN#PPS)	Ours (R2D2+NN @SIFT+NN#PPS)	Ours (SP+SG @SIFT+NN#PPS)
			
	D2-Net+NN @SIFT+NN	R2D2+NN @SIFT+NN	SP+SG @SIFT+NN



So from optimization of contrast, experiment shows that the precision PixSfM optimization method in 3D reconstruction precision public data sets South Building, British Museum, Sacre Coeur which can enhance the effect of 3D reconstruction, compared to the optimized ago, can significantly improve the accuracy of the three-dimensional reconstruction of point position.

6 SUMMARY

In this paper, we propose a multi-model fusion SfM framework called Mixum, which can fuse deep learning with traditional feature extraction and matching, and then eliminate redundant low-reliable duplicates and near-duplicates so as to make full use of the advantages of different algorithms and compensate for each other's shortcomings. Through this framework, more comprehensive and richer feature information can be obtained, and then the 3D reconstruction results can be improved. Then, the PixSfM accuracy optimization method is combined by the Mixum framework, and the PixSfM algorithm uses the depth features to optimize the error in the 3D reconstruction process. When the Mixum framework is combined with the PixSfM method, the accuracy of the reconstruction results can be further improved by the Mixum framework.

By selecting multiple public data sets as test data sources, the above two methods are compared. The experimental results show that the Mixum framework can effectively improve the 3D reconstruction results when the influence factors such as illumination, noise and representation of the data source change greatly. The 3D points are significantly increased, and the average reprojection error is reduced by 12%. When the PixSfM accuracy optimization algorithm is combined with the Mixum framework, the experimental results show that the average reprojection error index is reduced by 23%, which significantly improves the accuracy of 3D reconstruction points.

Future research can be carried out in the following ways.

(1) Although the multi-feature model fusion method proposed in this paper has achieved optimization results, the reconstruction time has increased a lot because the method has undergone multiple optimization processes, so it is limited to scenes with low requirements for reconstruction time. Further research can be done to solve the problem of long time consumption.

(2) This paper mainly studies the optimization of feature point extraction and matching stage in SfM 3D reconstruction, and there are still some stages that can be optimized in other stages of SfM 3D reconstruction, such as BA optimization problem and so on.

(3) The main idea of this paper is still to build on the original 3D reconstruction process and integrate deep learning for optimization. In fact, there have been many end-to-end solutions of deep learning, which directly omit the intermediate process for reconstruction, which is worthy of further research.

(4) In the point cloud results of the reconstruction experiment in this paper, it can be seen that due to the limited shooting Angle of most Internet images, most of them focus on the iconic front, so most of them can only reconstruct part. In the future, the method of artificial intelligence content generation can be considered to complete the remaining parts.

In summary, 3D reconstruction is a complex and challenging research field, and there are still many problems and directions worthy of in-depth research in the future. Through continuous exploration and innovation, it is believed that the accuracy and efficiency can be further improved, and more possibilities can be provided for the development and industrial landing of 3D reconstruction.

7 REFERENCES

[1] González Izard, S., Sánchez Torres, R., Alonso Plaza, O., Juanes Mendez, J. A., & García-Peñalvo, F. J. (2020). Nextmed: Automatic imaging segmentation, 3D reconstruction, and 3D model visualization platform using augmented and virtual reality. *Sensors*, *20*(10), 2962. https://doi.org/10.3390/s20102962

[2] Neves, M., Marques, B., Madeira, T., Dias, P., & Santos, B. S. (2022). Using 3D reconstruction to create pervasive augmented reality experiences: A comparison. *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 712-713. https://doi.org/10.1109/VRW55335.2022.00207

[3] Ma, Z. & Liu, S. (2018). A review of 3D reconstruction techniques in civil engineering and their applications. *Advanced Engineering Informatics*, *37*, 163-174. https://doi.org/10.1016/j.aei.2018.05.005

[4] Xu, Y. & Stilla, U. (2021). Toward building and civil infrastructure reconstruction from point clouds: A review on data and key techniques. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 2857-2885. https://doi.org/10.1109/JSTARS.2021.3060568

[5] Shao, J. (2021). Testing object detection for autonomous driving systems via 3d reconstruction. *2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, 117-119. https://doi.org/10.1109/ICSE-Companion52605.2021.00052

[6] Bevilacqua, M. G., Russo, M., Giordano, A., & Spallone, R. (2022). 3D reconstruction, digital twinning, and virtual reality: Architectural heritage applications. *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 92-96. https://doi.org/10.1109/VRW55335.2022.00031

[7] Maken, P. & Gupta, A. (2023). 2D-to-3D: A Review for Computational 3D Image Reconstruction from X-ray Images. *Archives of Computational Methods in Engineering*, *30*(1), 85-114. https://doi.org/10.1007/s11831-022-09790-z

[8] Hržica, M., Cupec, R., & Petrović, I. (2021). Active vision for 3D indoor scene reconstruction using a 3D camera on a pan-tilt mechanism. *Advanced Robotics*, *35*(3-4), 153-167.

https://doi.org/10.1080/01691864.2021.1875042
[9] Se, S. & Pears, N. (2020). Passive 3D Imaging. *3D Imaging, Analysis and Applications*, 39-107. https://doi.org/10.1007/978-3-030-44070-1_2
[10] Huang, Z. & Li, D. (2023). A 3D reconstruction method based on one-dimensional galvanometer laser scanning system. *Optics and Lasers in Engineering*, *170*, 107787. https://doi.org/10.1016/j.optlaseng.2023.107787
[11] Liu, L., Cai, H., Tian, M., Liu, D., Cheng, Y., & Yin, W. (2023). Research on 3D reconstruction technology based on laser measurement. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, *45*(6), 297. https://doi.org/10.1007/s40430-023-04231-9
[12] El Natour, G., Ait-Aider, O., Rouveure, R., Berry, F., & Faure, P. (2015). Toward 3D reconstruction of outdoor scenes using an MMW radar and a monocular vision sensor. *Sensors*, *15*(10), 25937-25967. https://doi.org/10.3390/s151025937
[13] Sun, Y., Huang, Z., Zhang, H., Cao, Z., & Xu, D. (2021). 3DRIMR: 3D reconstruction and imaging via mmWave radar based on deep learning. *2021 IEEE International Performance, Computing, and Communications Conference (IPCCC)*, 1-8. https://doi.org/10.1109/IPCCC51483.2021.9679394
[14] Marinello, F., Pezzuolo, A., Cillis, D., & Sartori, L. (2016). Kinect 3D reconstruction for quantification of grape bunches volume and mass. *Engineering for Rural Development*, *15*, 876-881.
[15] Valgma, L. (2016). *3D reconstruction using Kinect v2 camera*. University of Tartu.
[16] Zhang, D., Du, C., Peng, Y., Liu, J., Mohammed, S., & Calvi, A. (2024). A multi-source dynamic temporal point process model for train delay prediction. *IEEE Transactions on Intelligent Transportation Systems*. https://doi.org/10.1109/TITS.2024.3430031
[17] Zhou, K., Song, L., Hu, J., Guo, S., Dong, Y., Sun, Y., & Xu, Y. (2021). Real-time 3D reconstruction of dynamic scenes with multiple Kinect V2 sensors. *International Broadcast Convention*.
[18] Yang, G., Cui, Y., Belongie, S., & Hariharan, B. (2018). Learning single-view 3d reconstruction with limited pose supervision. *Proceedings of the European Conference on Computer Vision (ECCV)*, 86-101. https://doi.org/10.1007/978-3-030-01267-0_6
[19] Xu, Q., Wang, W., Ceylan, D., Mech, R., & Neumann, U. (2019). Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in Neural Information Processing Systems*, 32.
[20] Rebecq, H., Gallego, G., Mueggler, E., & Scaramuzza, D. (2018). EMVS: Event-Based Multi-View Stereo - 3D Reconstruction with an Event Camera in Real-Time. *International Journal of Computer Vision*, *126*(12), 1394-1414. https://doi.org/10.1007/s11263-017-1050-6
[21] Zhu, Q., Min, C., Wei, Z., Chen, Y., & Wang, G. (2021). Deep Learning for Multi-View Stereo via Plane Sweep: A Survey. No. arXiv:2106.15328.
[22] Kato, H. & Harada, T. (2019). Learning view priors for single-view 3d reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9778-9787. https://doi.org/10.1109/CVPR.2019.01001
[23] Yanwen, Z., Kai, H., & Pengsheng, W. (2020). Review of 3D reconstruction algorithms. *Nanjing Xinxi Gongcheng Daxue Xuebao*, *12*(5), 591-602.
[24] Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., & Szeliski, R. (2006). A comparison and evaluation of multi-view stereo reconstruction algorithms. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, *1*, 519-528. https://doi.org/10.1109/CVPR.2006.19
[25] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2022). NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, *65*(1), 99-106. https://doi.org/10.1145/3503250
[26] Fu, K., Peng, J., He, Q., & Zhang, H. (2021). Single image 3D object reconstruction based on deep learning: A review. *Multimedia Tools and Applications*, *80*(1), 463-498. https://doi.org/10.1007/s11042-020-09722-8
[27] Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., & Sattler, T. (2019). D2-net: A trainable cnn for joint description and detection of local features. *Proceedings of the IEEE/Cvf Conference on Computer Vision and Pattern Recognition*, 8092-8101. https://doi.org/10.1109/CVPR.2019.00828
[28] Revaud, J., De Souza, C., Humenberger, M., & Weinzaepfel, P. (2019). R2d2: Reliable and repeatable detector and descriptor. *Advances in Neural Information Processing Systems*, *32*.
[29] DeTone, D., Malisiewicz, T., & Rabinovich, A. (2018). Superpoint: Self-supervised interest point detection and description. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 224-236. https://doi.org/10.1109/CVPRW.2018.00060
[30] Sarlin, P.-E., DeTone, D., Malisiewicz, T., & Rabinovich, A. (2020). Superglue: Learning feature matching with graph neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4938-4947. https://doi.org/10.1109/CVPR42600.2020.00499
[31] Sun, J., Shen, Z., Wang, Y., Bao, H., & Zhou, X. (2021). LoFTR: Detector-free local feature matching with transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8922-8931. https://doi.org/10.1109/CVPR46437.2021.00881
[32] Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., & Brox, T. (2017). Demon: Depth and motion network for learning monocular stereo. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5038-5047. https://doi.org/10.1109/CVPR.2017.596
[33] Tang, C. & Tan, P. (2019). BA-Net: Dense bundle adjustment networks. *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*.
[34] Lindenberger, P., Sarlin, P.-E., Larsson, V., & Pollefeys, M. (2021). Pixel-perfect structure-from-motion with featuremetric refinement. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5987-5997. https://doi.org/10.1109/ICCV48922.2021.00593

Contact information:

Dong LI
China Railway Smart City Research and Development Center,
China Railway Liuyuan Group Co., Ltd.
No. 36 Zhonghuan West Road, Tianjin Airport Economic Area, Tianjin, China
E-mail: bj1858@163.com

Gongyun FU
China Railway Smart City Research and Development Center,
China Railway Liuyuan Group Co., Ltd.
No. 36 Zhonghuan West Road, Tianjin Airport Economic Area, Tianjin, China
E-mail: fgy@outlook.com

Shunjie YANG
(Corresponding author)
School of Software Engineering,
Beijing Jiaotong University,
No. 3 Shangyuancun, Haidian District, Beijing, China
E-mail: sj1136455648@163.com