International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 3, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v3n1.13
Journal homepage: https://wepub.org/index.php/IJCSIT/index

A Review of Traffic Scene Reconstruction Based on Images and Point Clouds

Xiaoning Dong

School of Information engineering, Henan University of Science and Technology, Luo Yang 477100, China

ABSTRACT

This paper elaborates on a three-dimensional scene reconstruction method based on point clouds, images, and the fusion of images and point clouds. Relevant evaluation indicators are used to evaluate the performance of traffic scene reconstruction technology. The problems in traffic reconstruction under image and point cloud elements are analyzed and summarized. Finally, the challenges and future research directions in the field of traffic scene reconstruction based on images and point clouds are pointed out.

KEYWORDS

Traffic scenarios; 3D reconstruction; Image point cloud fusion; Multi element data; Semantic segmentation

1. INTRODUCTION

Traffic scene reconstruction refers to the use of computer vision and computer graphics technology to digitize real traffic scenes and generate realistic 3D models. Traffic scene reconstruction is an interdisciplinary field that combines sensor data processing with computer graphics to generate three-dimensional models with strong realism and high accuracy [1]. The reconstruction of traffic scenes can help solve many related problems in various fields:

1) Transportation Planning: Utilizing 3D model visualization to better understand the city's transportation system, such as roads, traffic lights, pedestrian overpasses, underground garages, etc., to provide assistance for urban transportation planning.

2) Road design: Using 3D models can simulate and evaluate road design schemes more realistically, including road width, number of lanes, traffic signal lights, and optimization in traffic congestion and safety issues.

3) Autonomous driving: Using 3D models, real-time perception of traffic scenes can be achieved to assist autonomous vehicles in path planning, obstacle avoidance, and environmental perception functions.

3) Traffic safety: Through digital modeling of traffic scenes, it is possible to better analyze the causes of traffic accidents and predict potential traffic safety hazards, thereby improving the level of traffic safety.

5) Games and Virtual Reality: Traffic scene reconstruction can also be applied in fields such as games and virtual reality, providing users with more realistic games and experiences.

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

In summary, application in transportation scenarios with multi factor data can help transportation systems achieve comprehensive digital transformation and upgrading, improve the safety, reliability, and efficiency of transportation systems, and contribute to the sustainable development of urban transportation [2].

This article analyzes and introduces the research methods in existing literature, summarizes the mainstream 3D reconstruction methods used for multi element data traffic scene reconstruction, compares and analyzes them, summarizes the advantages and disadvantages of various reconstruction methods, and finally looks forward to and summarizes the application prospects of traffic scenes.

2. OVERVIEW OF RELEVANT DEFINITIONS

2.1. Concept and Key Technologies of Traffic Scene Construction

When it is necessary to establish a digital model of a real object or system and conduct simulation analysis on it, if the accuracy of this digital model is not high, the simulation results will lose reference value. 3D reconstruction refers to the process of transforming real-world objects or scenes into digital models through certain technical means (such as optical scanning, photogrammetry, LiDAR, etc.), which can be used for visualization, simulation, virtual reality, and other aspects. Therefore, 3D reconstruction can provide an accurate digital model for digital twins. Traffic scene reconstruction refers to the process of using sensor data (such as cameras, LiDAR, etc.) and computer vision technology to 3D model and reconstruct vehicles, pedestrians, roads, traffic signals, etc. in traffic scenes.

The key technologies for scene reconstruction include the following aspects:

1) Sensor data collection: Using sensors such as cameras and LiDAR to collect data on traffic scenes, obtaining information on vehicles, pedestrians, roads, buildings, and other information in the scene.

2) Data preprocessing: Preprocessing the collected data, including data denoising, point cloud registration, point cloud segmentation, etc., to improve the quality and accuracy of the data.

3) Object detection: Using computer vision technology to detect objects in the scene, including vehicles, pedestrians, buildings, as well as their position, posture, and other information in three-dimensional space.

4) Map construction: Integrate the detected object information, as well as information such as roads and traffic lights, to construct a three-dimensional map, including road topology, lane lines, traffic lights, etc.

5) Scene reconstruction: In the constructed 3D map, 3D reconstruction is performed on objects such as vehicles and pedestrians to provide more accurate position, posture, and other information, supporting application scenarios such as vehicle autonomous navigation and traffic congestion monitoring.

At present, 3D reconstruction methods can be divided into two categories based on the use of sensors:

1) Using active distance sensors, such as laser scanners, structured light, etc;

2) Relying on passive distance sensors, such as cameras.

2.2. Classification of Traffic Scene Elements and Graph Models

In traffic scenes, traffic elements are divided into two categories: background elements and foreground elements. There are a large number of background elements that constitute traffic scenes, such as the sky, road surface, and the sides of walls. The foreground elements include traffic lights, traffic signs, vehicles, pedestrians, etc. Classify each element and define it uniformly using a graph

model [3]. For example, for the i-th image, Gi represents traffic scene elements, BS_i represents background scenes, and FS_i represents foreground scenes. We define the traffic scene of the i-th image as follows:

$$G_i = \{BS_i, FS_i\} \tag{1}$$

The background scenes include "sky", "sidewalk", "left wall", "right wall", and "background wall". The background elements are represented as a set:

$$BS_i = \{SK_i, RS_i, LW_i, RW_i, BW_i\} \tag{2}$$

Among them, SK_i, RS_i, LW_i, RW_i, and BW_i represent the "sky", "road surface", "left wall", "right wall", and "background wall" of the first image, respectively.

The prospect is FS_i, consisting of traffic lights, traffic signs, and moving objects:

$$FS_i = \{L_i, TS_i, MO_i\} \tag{3}$$

Among them, L_i represents traffic lights, TS_i represents traffic signs, and MO_i represents moving objects, namely pedestrians and moving vehicles.

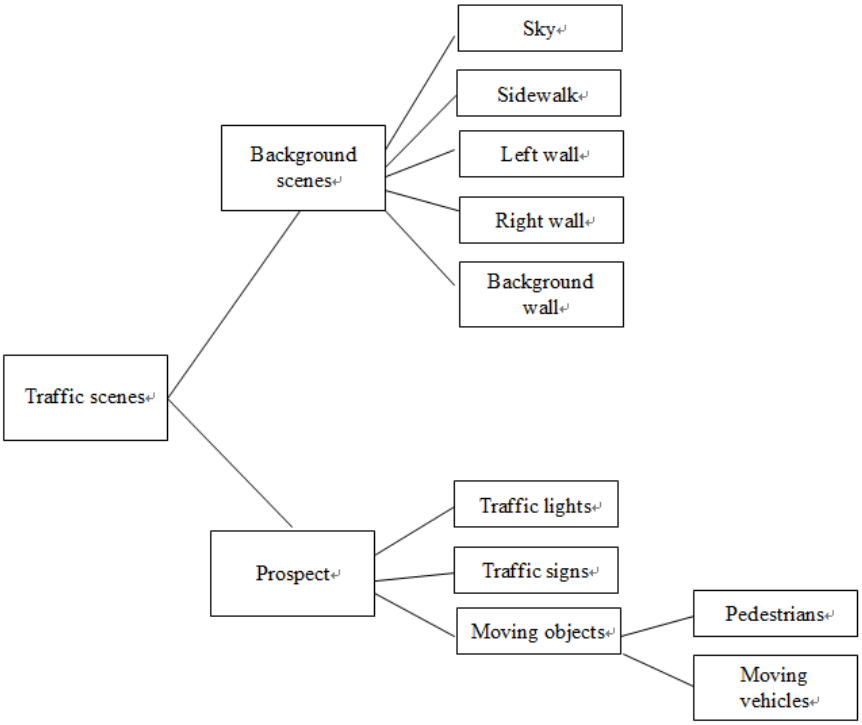

Figure 1. Graph model of the traffic scenes

3. THREE DIMENSIONAL RECONSTRUCTION METHODS FOR TRAFFIC SCENES

In the field of transportation, 3D scene modeling and digital twins can be combined to achieve visualization and simulation of traffic scenes. Specifically, three-dimensional scene modeling technology can be used to construct three-dimensional models of real traffic scenes, including roads, buildings, vehicles, pedestrians, etc., and digital twin technology can be used to copy them into computers to achieve virtual simulation and visualization of traffic scenes [4]. This can provide convenient analysis tools for traffic planning, traffic safety, and other fields, such as simulating and

analyzing traffic flow, traffic accidents, etc. In addition, digital twins can also reflect real-time changes in real-life scenarios into digital twin models through real-time data updates, making it easier to simulate and analyze more accurately. Therefore, the combination of 3D scene modeling and digital twins can achieve comprehensive, realistic, and reliable virtual simulation of traffic scenes, provide data support for decision-making in the transportation field, and improve the efficiency and safety of transportation.

This article provides a classification overview of 3D scene reconstruction methods based on different data sources, including point cloud based 3D reconstruction, image based 3D reconstruction, and image point cloud fusion based 3D reconstruction. This section summarizes and analyzes several commonly used method models, compares the advantages, disadvantages, and reconstruction effects of each method [5].

3.1. 3D Scene Reconstruction Based On Laser Point Cloud

Lidar (Light Detection and Ranging, LiDAR) has the characteristics of fast data acquisition speed, minimal environmental impact, direct acquisition of three-dimensional coordinates of target objects, and high accuracy. It has become the mainstream means of obtaining three-dimensional data information. It uses a laser rangefinder and optical mechanical structure to scan the surrounding environment, determines the spatial distance between the laser and the target object by recording the flight delay of the laser pulse, and determines the pulse emission angle by recording the rotation angle of the optical mechanical structure. When a navigation system provides precise position and attitude information, the point cloud in local coordinates can be transformed into the world coordinate system through distance and angle, thereby obtaining the precise position coordinates of the external environment in the world coordinate system.

3.1.1. Reconstruction method based on geometric features

This type of method can be divided into the following types:

1) Reconstruction method based on plane fitting.

This method mainly generates a three-dimensional model of planar objects in the scene by plane fitting point cloud data. The commonly used algorithm is the Random Sample Consensus (RANSAC) algorithm, but it can only segment a single plane. Negri et al. [6] proposed the Ground Plane Fitting (GPF) algorithm, which screens the fitting points and accelerates the fitting speed, but the accuracy of ground segmentation is not high. Cui et al. [7] divided the scene, filtered dynamic regions, and then used the GPF algorithm to fit the ground plane, proposing the R-GPF algorithm. This method improves the accuracy of ground segmentation compared to the GPF algorithm, but is more sensitive to outliers. In the same year, the author proposed a ground segmentation algorithm called Patchwork, which is based on the concentric region model to represent point clouds, uses GPF to fit the ground, and then uses the direction, height, and flatness information of the ground normal vector to segment the ground. The experimental results show that the F1 score of this algorithm can reach 93%, and the operating frequency exceeds 30Hz. However, ground point clouds are prone to oversegmentation, and such methods usually divide the scene point cloud into multiple small regions, requiring ground segmentation of multiple regions, resulting in slower speeds. The advantage of this type of method is that it can accurately reconstruct planar objects, but the reconstruction effect is poor for non-planar objects [8].

2) Reconstruction method based on curvature estimation.

This method mainly estimates the curvature of point cloud data to identify and segment objects with different curvatures, and generates their one-dimensional models. Since the 1880s, Kacnderink was the first to introduce the concept of differential geometry into computer vision and proposed the concept of curvature estimation. Domestic and foreign scholars began systematic research on point cloud curvature estimation methods, and successively proposed various curvature estimation methods

such as local surface fitting methods and discrete equation approximation methods. Local surface fitting refers to the use of quadratic surfaces, polynomial surfaces, or other methods to first mesh the surface, and then calculate the curvature by calculating the angle between the neighboring polygons of each mesh vertex. Wang et al. [9] proposed a method based on Euler's theorem, which then solved curvature through least squares; Martin proposed the parabolic fitting method. The basic idea of this method is to use a small parabolic surface to fit various points and their neighboring points on a local surface according to the principle of least squares. After fitting the surface, the curvature and direction of the point can be calculated by solving the equation of the parabolic surface; Lim et al. [10] calculated point cloud curvature by performing least squares fitting on all discrete normal curvatures corresponding to adjacent points. The advantage of this type of method is that it can accurately reconstruct objects with surfaces, but the reconstruction effect is poor for planar objects.

3.1.2. Reconstruction methods based on deep learning

The method of voxelizing unordered point clouds into regular 3D meshes is an alternative to processing point clouds to adapt to deep neural networks. For example, Schmohl S and others first voxelize ALS point clouds and then apply them to submanifold sparse convolutional networks for processing. However, voxelization inevitably leads to information loss and generates artifacts, which have a negative impact on the learning of 3D features. In addition, the large number of unoccupied grids stored in voxel structures will result in high memory requirements. Some researchers have also attempted to directly apply convolutional operators to the original point cloud and use deep neural networks to learn advanced point features. For example, Lim et al. [11] proposed a fully convolutional network that takes the original coordinates of the input point cloud and three additional spectral features extracted from geographical reference images at the same location as inputs for point by point classification; WANG S and others have developed multi-scale deep neural networks to achieve more powerful feature learning and further improve point cloud classification performance. These methods first utilize a shared MLP network to extract features from each point; then, use downsampling blocks to aggregate the features of each point into cluster based features; finally, another MLP network is used followed by a Softmax classifier for point by point classification. Schmohl [13] et al. [12] proposed a direction constrained convolutional operator for point feature extraction and designed a multi-scale fully convolutional network for point cloud classification; Arief H A et al [14] developed the Atrous XCRF module to enhance the original PointCNN model and developed beneficial performance in airborne LiDAR point cloud classification. In the field of forestry, WEN C et al [15] [13] proposed a 3D point cloud semantic segmentation method based on KNN search, which improved the shortcomings of existing methods in local point cloud feature extraction and effectively improved the accuracy of semantic segmentation. Although these point cloud based methods have achieved excellent results in airborne laser point cloud classification, they cannot fully recognize fine-grained local structures due to the uneven density distribution of point cloud data.

3.2. Image Based Traffic Scene Reconstruction

The data obtained directly from two-dimensional images will lose some information compared to the real three-dimensional scene, so how to use two-dimensional images to reconstruct three-dimensional scenes has become a hot topic. This article divides current research methods into two categories based on different methods of reconstructing scenes:

1) Image-based scene layout extraction methods;

2) A scene reconstruction method based on image segmentation.

3.2.1. Image based scene layout extraction method

In order to understand the relationship between scene depth and image features, Arief et al. [14] used a Markov network in 2008 to infer the three-dimensional position and direction of each image patch

based on assumptions such as connectivity and coplanarity. This method has a good processing effect on outdoor scene images. One drawback of this method is that it requires the use of three-dimensional depth ground truth markers during training, which limits its applicability. Wang et al. [15] provided a multi hypothesis framework for reliable estimation of scene structure based on a single image. According to Hoiem's method, each image region is divided into three categories: "support", "vertical", and "sky". By modeling powerful clues of color and texture, objects corresponding to specific geometric classes can be implicitly identified. However, this method requires pixel level ground truth training data, which requires tedious labeling work. Lou i et al. [16] have many advantages in the scene stage. In scene reconstruction, the Scene Stage usually refers to the preliminary stage of scene modeling, also known as geometric reconstruction. The stage can provide the audience with information about semantic background, identity of scene elements, and so on. They proposed a method to infer weak scene sets from a single image and established a rough sketch of the scene model. Yuan et al. [17] proposed a framework for inferring pixel level 3D layouts through global image structures. They used image segmentation to obtain stage classification and generated 3D layouts through stage classification. Firstly, predict the global image structure, and then use the global structure for fine-grained pixel level 3D layout extraction. After obtaining the image level structure, use it as prior knowledge to infer pixel level 3D layout.

Zhu et al. [18] and Wang et al. [19] exported parallel lines in indoor scenes to infer a set of predefined models for generating 3D layouts. Barinova and others focus on urban scenes. By detecting parallel lines, vanishing points can be estimated and used to infer the 3D layout in the conditional random field framework. However, these algorithms are highly dependent on parallel lines and can only process indoor images or outdoor images containing parallel lines.

3.2.2. Scene reconstruction method based on semantic segmentation

The concept of semantic segmentation was first proposed by Lee et al. [20] which is defined as assigning a pre-defined label to each pixel in an image to represent its semantic category. Being able to extract the information that the image itself needs to express base on its texture, scene, and other high-level semantic features is more practical. It not only classifies each pixel in the image, but also annotates the object category that the pixel belongs to in the image. This not only segments the region, but also annotates the content of the region. Segmentation methods can be divided into traditional segmentation methods and deep network-based methods.

FCN is an improvement on the VGG-16 network architecture, which can restore the category of pixels during the segmentation process, greatly promoting the development of semantic segmentation. However, there are still two problems in this field: first, after the image is pooled, the spatial position information of some pixels is lost [21]; the second issue is that the segmentation process fails to effectively consider image context information, resulting in an imbalance in the utilization of local and global features. In response to these two issues, researchers have proposed a series of new methods based on FCN [22], which can be roughly divided into five categories according to their improvement characteristics: FCN based methods, optimized convolutional structure based methods, encoder decoder based methods, feature fusion based methods, and RNN based methods.

1) FCN based methods represent algorithms DeepLab, DeepLab-V2, DeepLab-V3, and CRFasRNN. Method features: Obtaining multi-scale image information through techniques such as image pyramid, perforated convolution, and perforated spatial pyramid pooling, with high invariance to spatial transformations. FCN has been optimized and improved to extract dense image features and increase receptive fields, and to use conditional random fields for structural prediction. Advantages: Improvements are made to address the shortcomings of FCN, which can effectively enhance the filter's field of view, obtain multi-scale representations of images, and improve the spatial accuracy of segmentation results. Disadvantages: Slow segmentation speed and unclear segmentation effect on small-scale objects.

2) Based on the method of optimizing convolutional structures, representative algorithms include Dilation10, DUC+HDC, and deformable. Advantages: Enlarge the receptive field, effectively slow down the speed of feature map resolution reduction, and save the spatial position information of pixels. Disadvantages: The continuity of local pixel information is disrupted, and the adaptability to unknown deformations is poor.

3) Based on encoding and decoding methods, representative algorithms include BayesianSegNet, DeconvNet, ENet, GCN+ [23]. The method features a decoder composed of deconvolution or upsampling operations to upsample low resolution feature maps. Advantages: Restoring the spatial dimension and pixel position information of the image, avoiding the problem of reduced feature map resolution after pooling operation. Disadvantages: Compared with FCN, its network has too many training parameters and a large computational workload;

4) Based on feature fusion methods, representative algorithms include LRR, RefineNet, PSPNet, ICNet, LC. Method features: By using cross layer structures, spatial pyramids, pooling modules, multi-scale convolutions, and cascading models, feature information of different scales and positions in the image is captured and fused, gradually refining the segmentation results. Advantages: Capturing contextual information and avoiding issues such as high computational complexity and memory consumption caused by using probability graph models. Disadvantage: Partial loss of boundary information for segmentation targets.

5) RNN based methods represent algorithms R-CNN, ReSeg, DAG-GNN, 2D-LSTM. Method features: The RNN [24] part is combined with convolutional layers and embedded into DNN. Local spatial features are extracted using convolutional layers in CNN, and pixel sequence features are extracted using RNN layers. The related network structures include ReNet, LSTM, and GRU. Advantages: It can recursively process historical information and model historical memory, making it easy to extract pixel sequence information and capture contextual information in images. Disadvantage: During the process of image pixel sequence, some pixel information may be lost.

3.3. 3D Scene Reconstruction Based On Image And Point Cloud Fusion

The disadvantage of point cloud reconstruction is that it has a high data density, high processing complexity, and requires a large amount of computing resources. In addition, point cloud data usually only provides geometric information of the scene, and cannot provide texture and color information of the scene [25]. Therefore, using point clouds alone for scene reconstruction may result in incomplete and accurate scene information. The disadvantage of image reconstruction is that it requires high requirements for the scene, requiring the shooting of high-quality images with multiple perspectives and consistent lighting conditions. In addition, image reconstruction usually only provides texture and color information of the scene, and cannot provide geometric information of the scene. Therefore, using images alone for scene reconstruction may result in inaccurate geometric information of the scene. In summary, there are some drawbacks to using point clouds or images alone for scene reconstruction. Therefore, in practical applications, it is usually necessary to combine the information of both and use image and point cloud fusion technology for scene reconstruction to obtain more comprehensive and accurate scene information [26]. To fuse point cloud and image data, it is necessary to perform pose estimation between two sensors, which is called external parameter calibration. In external parameter calibration, it is often necessary to establish connections in point cloud image data through geometric constraints, such as constraints between points, lines, and surfaces.

3.3.1. Method based on point-to-point constraints

Gupta et al. [27] designed an inverted triangle calibration object that calculates the coordinates of feature points in the calibration board coordinate system through a three-step interpolation fitting method. However, the results of this method are related to the accuracy of the calibration object, and solving point cloud coordinates with known geometric constraints often has low credibility. In the

method proposed by Badrinarayanan et al. [28] only one calibration board image needs to be collected, and then the point cloud is projected onto the image so that it coincides exactly with the rectangle in the image. The solution is solved through line line constraints and point point constraints, which have multiple parameters and often cannot guarantee accuracy due to the dispersion of points. The methods based on point-to-point constraints often have significant selection errors due to the low visibility of landmark points in the point cloud, resulting in results containing significant uncertainty.

3.3.2. Method based on point line constraints

Noh et al. [29] used a triangular calibration object to solve the external parameters of two-dimensional LiDAR by minimizing the distance from the point to the boundary and distance invariance. The results were good, but did not take into account the uneven distribution of the point cloud in the scanning plane. Zhao et al. [30] achieved better boundary endpoint extraction results by adding reflection bands at the edges, and the flatness of the pasted material directly affects the accuracy of the received point cloud. Zhao et al. [31] used a nonlinear optimization least squares algorithm to minimize the distance from the scanning point to the straight line formed by the scanning trajectory and calibration plane. Due to the sparsity of laser scanning points, the endpoints of point line constrained methods are often not accurate enough, which can easily lead to incorrect estimation.

3.3.3. Method based on point surface constraints

Chen et al. [32] determined the size of the calibrated cube box by minimizing the distance from the point to the plane. In order to reduce the error in extracting calibration object boundaries, Dai et al. [33] used the entire target point cloud when estimating vertices, thus avoiding the uncertainty of identifying edge points. The core principle is still based on point plane constraints. Ghiasi et al. [34] first proposed the use of chessboard grid calibration, which typically requires at least five different perspectives to observe by solving the constraints from the scanning point to the calibration board plane. On this basis, Lin et al. [35] provided the minimum solution for this method. Point surface constraints often need to be combined with other constraints, especially when the density of the scanned calibration board point cloud is uneven or there is a lot of surface noise, the accuracy of obtaining the plane normal vector is crucial.

3.3.4. Method based on face face constraints

Li et al. [36] used a V-shaped calibration board to paste two checkerboard patterns onto two calibration board planes at a certain angle. They constrained the normal vectors in the image and point cloud by solving for them, and also added the reverse projection plane of the calibration board boundary as a constraint in the vertical direction, which achieved good results. Similar constraints include Pinheiro et al. [37], which are based on normal vectors being parallel to each other and boundary reverse projection plane being perpendicular to the normal direction. Shuai et al. [38] established a three-dimensional Cartesian coordinate system at the corners of building walls and constrained it by projecting the normal vector of the plane perpendicular to the intersection line between the three planes, but did not take into account the errors caused by uneven wall corners. The method based on face to face constraints can often quickly determine the rotation matrix under conditions of limited data, but the determination of the translation vector still requires additional constraints.

4. COMMON DATASETS AND PERFORMANCE EVALUATION

4.1. Common Datasets

The following are some commonly used datasets for traffic scene reconstruction:

1) KITTI dataset: The KITTI dataset is a widely used computer vision dataset used for tasks such as autonomous driving and 3D scene reconstruction. This dataset was jointly created by Karlsruhe

Institute of Technology and Toyota Motor Corporation, and includes high-resolution images from cars, LiDAR point clouds, and 3D bounding boxes. The KITTI dataset contains the following subsets:

Object detection: This subset contains a large amount of images, LiDAR data, and 3D object annotation, used for tasks such as object detection and tracking.

Target tracking: This subset contains a series of image and object tracking data for tasks such as target tracking and multi-target tracking.

Semantic segmentation: This subset contains high-resolution urban scene images for tasks such as scene segmentation and semantic segmentation.

Road/lane detection: This subset includes images from cars and LiDAR data for tasks such as road and lane detection.

The KITTI dataset is a very useful dataset for training and evaluating algorithms and models for tasks such as autonomous driving and 3D scene reconstruction. Meanwhile, this dataset has also become an important benchmark testing dataset in the field of computer vision.

2) The Cityscapes dataset: The Cityscapes dataset is a high-quality dataset used for tasks such as urban scene understanding and autonomous driving. This dataset was created by a computer vision team from the University of Stuttgart in Germany and includes high-resolution images, semantic segmentation annotations, and instance segmentation annotations from different cities. The Cityscapes dataset contains 5000 high-resolution images from 50 different cities including Germany, France, and Switzerland. For each image, pixel level semantic segmentation annotations are provided, including 19 different categories such as roads, buildings, vehicles, pedestrians, etc. For some images, instance level segmentation annotations are also provided to annotate information such as bounding boxes and instance IDs for each object. The Cityscapes dataset is a very useful dataset for training and evaluating algorithms and models for tasks such as urban scene segmentation and autonomous driving. Meanwhile, this dataset has also become an important benchmark testing dataset in the field of computer vision.

3) The Apollo Scape dataset: It is a dataset provided by the Apollo Open Platform, which includes various traffic scenarios such as urban roads, highways, tunnels, etc. It provides images, LiDAR data, and 3D object annotation. This dataset was jointly created by the Apollo team and the Beijing Automotive Engineering Research Institute, and is an open dataset that can be downloaded and used for free. The Apollo Scape dataset contains the following data: 380000 images, covering various traffic scenarios such as urban roads, highways, tunnels, etc; 3D object annotation data, including 3D bounding boxes and trajectory information of objects such as vehicles, pedestrians, bicycles, etc; Annotate scene segmentation at the pixel level for each image.

4) The nuScenes dataset: It is a dataset for autonomous driving that includes high-quality images, LiDAR data, and 3D object annotation from cities such as New York, Boston, and Singapore. The nuScenes dataset is a high-quality dataset for autonomous driving, which includes high-resolution images, LiDAR data, and 3D object annotation from cities such as New York, Boston, and Singapore. This dataset was created by nuTonomy company and later acquired by Uber for open source, making it available for free download and use. The nuScenes dataset contains the following data:

Image data: including image data for 1000 scenes, covering various traffic scenarios such as urban roads and highways.

Lidar data: including 20 frames per second of lidar data, used to construct a 3D point cloud map.

3D object annotation: including 3D bounding boxes and trajectory information of objects such as vehicles, pedestrians, bicycles, etc.

Scene semantic segmentation: including the pixel level scene segmentation annotation of each image.

5) Waymo Open dataset: It is the dataset opened by Waymo, an autonomous driving company under Google, and contains high-resolution images, laser radar data and 3D object annotation from autonomous vehicle. The Waymo Open dataset is a large-scale autonomous driving dataset released by Google's autonomous driving project Waymo, which includes high-resolution images, LiDAR data, GPS trajectories, and object annotation data. This dataset aims to provide researchers and developers with a public resource for researching autonomous driving technology. Includes the following content:

Image data: including high-resolution front camera and side camera images, as well as panoramic images captured by car mounted cameras.

Lidar data: including up to 100000 points per second of lidar data, used to construct three-dimensional point cloud maps.

GPS trajectory: including global positioning system (GPS) trajectory data of the vehicle.

Object annotation: includes three-dimensional bounding boxes and trajectory information of objects such as vehicles, pedestrians, bicycles, etc.

4.2. Evaluation indicators

1) Intersection over Union (IoU): Used to evaluate the quality of output results. The threshold for IoU is set to 0.3, and the predicted probability output is binarized and IoU is calculated. Its definition is shown in formula 4. In the context of 3D reconstruction, IoU refers to the ratio of the intersection of the predicted 3D shape volume and the actual ground volume to the union of the two volumes. The higher the IoU value, the higher the quality of the reconstruction result.

$$IoU(G,R)=\frac{|G\cap R|}{|G\cup R|} \tag{4}$$

Among them, G and R represent the occupancy maps of binary classification.

2) Chamfer distance (CD): defines the chamfer distance between the real shape G and the reconstructed shape R (both represented as point clouds), which calculates the average shortest distance between the generated G and the original R. In 3D reconstruction, the smaller the value of CD, the better.

$$CD(G,R)=\frac{1}{|R|}\sum_{r\in R}\min_{g\in G}\left\|r-g\right\|_2+\frac{1}{|G|}\sum_{g\in G}\min_{r\in R}\left\|g-r\right\|_2 \tag{5}$$

3) Earth Mover's Distance (EMD): used to indicate the degree of approximation between two distributions S1 and S2.

$$d_{EMD}(S_1,S_2)=\min_{\Phi:S_1\to S_2}\sum_{x\in S_1}\left\|x-\Phi(x)\right\|_2 \tag{6}$$

4) F-Score refers to the harmonic average of reconstruction accuracy and recall.

$$F(\tau)=\frac{2P(\tau)\ R(\tau)}{P(\tau)+R(\tau)} \tag{7}$$

5. FUTURE CHALLENGES AND PROSPECTS

By conducting three-dimensional modeling of the traffic environment and constructing a digital twin system, traffic conditions can be comprehensively and in real-time controlled, while providing basic

recognition data for the upcoming era of autonomous driving. After years of technological evolution, there are still many challenges in the construction, fusion, and visualization of 3D modeling methods for twin traffic environment data

(1) Difficulty in multi-source data fusion

In the construction of digital twin scenes in transportation, the format of 3D models constructed for different scales is not uniform, and models need to be aggregated, combined, and arranged, often requiring format conversion and adaptation. On the one hand, during the process of format conversion, information loss and distortion are often faced; On the other hand, the fusion display platform has high adaptation costs for multi-scale models, requires compatibility with different accuracies, and ensures unified visualization effects, which is difficult.

(2) Difficulty in large-scale structured modeling

Digital twin large-scale structured modeling refers to the digital modeling of large-scale structures (such as large buildings, roads, vehicles, etc.) to achieve virtual testing, optimized design, real-time monitoring, and other purposes. The modeling process of digital twins requires consideration of many details, such as fine modeling of physical properties, nonlinear response of structures, and so on. Therefore, extensive calibration and validation are required when constructing digital twin models.

(3) Difficulty in updating models quickly

The changes in the natural environment and the activities of human society have led to continuous changes in the transportation environment. The digital twin space should be able to quickly map changes and make updates and adjustments. However, the current 3D modeling methods and techniques mostly adopt the technical route of data collection, data processing, model optimization, and data publishing, which often requires a lot of post work and cannot adapt to the rapid changes and real-time updates of the model.

In order to address the aforementioned difficulties and challenges, a new type of transportation environment surveying and mapping will inevitably be developed in the future under the trend of digital transformation. Firstly, utilizing modeling equipment and methods with higher levels of automation and intelligence, quickly construct holographic digital models covering two and three dimensions. Secondly, modeling achievements will develop towards a structured and semantic direction, achieving human comprehensibility and machine recognizability. Once again, by utilizing various technologies such as mobile measurement and crowdsourcing mapping, we can perceive and update changes in the transportation environment in a timely manner, thereby improving the current status of the model results.

REFERENCES

[1] Li Y, Zhu C, Liu Y, et al. Geometric and semantic analysis of road image sequences for traffic scene construction [J]. Neurocomputing, 2021, 365: 336-339.

[2] Tao F, Zhang H, Liu A, et al. Digital Twin in Industry: State-of-the-art [J]. IEEE Transactions on Industrial Informatics (S0956-5515), 2018, 15(3): 2305-2315.

[3] Gomez-Ojeda R, Briales J, Fernandez-Moral E, et al. Extrinsic calibration of a 2d laser-rangefinder and a camera based on scene corners [C], in : IEEE International Conference on Robotics and Automation. IEEE, 2015, 3611-3616.

[4] Babahajiani P, Fan L, Kamarainen J K, et al. Comprehensive automated 3D urban environment modelling using terrestrial laser scanning point cloud [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016: 10-18.

[5] Wang J, Ye L, Gao R X, et al. Digital Twin for rotating machinery fault diagnosis in smart manufacturing [J]. International Journal of Production Research, 2019, 57(12): 3920-3933.

[6] Negri E, Fumagalli L, Macchi M. A review of the roles of digital twin in CPS-based production systems [J]. Procedia manufacturing, 2017, 11: 939-938.

[7] Cui Y, Chen R, Chu W, et al. Deep learning for image and point cloud fusion in autonomous driving: A review [J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(2): 722-739.

[8] Zhou L, Li Z, Kaess M. Automatic extrinsic calibration of a camera and a 3D lidar using line and plane correspondences [C], in : IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2018, 5562-5569.

[9] Wang S, Zhang F, Qin T. Research on the construction of highway traffic digital twin system based on 3D GIS technology [C]//Journal of Physics: Conference Series. IOP Publishing, 2021, 1802(3): 032035.

[10] Lim H, Hwang S, Myung H. ERASOR: egocentric ratio of pseudo occupancy-based dynamic object removal for static 3D Point cloud map building [J]. IEEE Robotics and Automation Letters, 2021, 6(2):2272-2279.

[11] Lim H, Oh M, Myung H. PatchWork: concentric zone-based region-wiseground segmentation with ground likelihood estimation using a 3D LiDAR sensor [J]. IEEE Robotics and Automation Letters, 2021, 6(3):6358-6365.

[12] Schmohl S, Sörgel U. Submanifold sparse convolutional networks for semantic segmentation of large-scale ALS point clouds [J]. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2019, 3: 77-83.

[13] WEN C, YANG L, LI X, et al. Directionally Constrained Fully Convolutional Neural Network for Airborne LiDAR Point Cloud Classification [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020(162):50-6.

[14] Arief H A, Indahl U G, Strand G H, et al. Addressing Overfitting on Point Cloud Classification Using Atrous XCRF [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2019(155):90-101.

[15] WANG Y J, JIANG T P, LIU J, et al. Hierarchical Instance Recognition of Individual Roadside Trees in Environmentally Complex Urban Areas from UAV Laser Scanning Point Clouds [J]. ISPRS International Journal of Geo-information, 2020(9):595.

[16] Lou Z, Gevers T, Hu N. Extracting 3D layout from a single image using global image structures [J]. IEEE Transactions on Image Processing, 2015, 24(10): 3098-3108.

[17] Yuan J, Li Y, Pan H, et al. 3D traffic scenes construction and simulation based on scene stages [C]//2018 Chinese Automation Congress (CAC). IEEE, 2018: 1334-1339.

[18] Zhu C, Li Y, Liu Y, et al. Road scene layout reconstruction based on CNN and its application in traffic simulation [C]//2019 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2019: 480-485.

[19] Wang H, Gould S, Roller D. Discriminative learning with latent variables for cluttered indoor scene understanding [J]. Communications of the ACM, 2013, 56(4): 92-99.

[20] Lee S, Huh S, Yoo D, et al. Rich feature hierarchies from omni-directional RGB-DI information for pedestrian detection [C]. 2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), 2015: 362-367.

[21] Byeon W, Breuel T M, Raue F, et al. Scene labeling with lstm recurrent neural networks [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3547-3555.

[22] Cai H, Pang W, Chen X, et al. A novel calibration board and experiments for 3D lidar and camera calibration. Sensors [J], 2020, 20(4): 1130.

[23] Liao Q, Chen Z, Liu Y, et al. Extrinsic calibration of lidar and camera with polygon [C], in: IEEE International Conference on Robotics and Biomimetics. IEEE, 2018, 200-205.

[24] Ye Q, Shu L, Zhang W. Extrinsic calibration of a monocular camera and a single line scanning lidar [C], in : IEEE International Conference on Mechatronics and Automation. IEEE, 2019, 1047-1054.

[25] Huang J K, Grizzle J W. Improvements to target-based 3D lidar to camera calibration. IEEE Access [J], 2020, 8: 134101-134110.

[26] Dong W, Isler V. A novel method for the extrinsic calibration of a 2-D laser-rangefinder & a camera [C], in: IEEE International Conference on Robotics and Automation. IEEE, 2017, 5104-5109.

[27] Gupta S, Girshick R, Arbeláez P, et al. Learning rich features from RGB-D images for object detection and segmentation [C]. European conference on computer vision, 2014: 345-360.

[28] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.

[29] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation [C]. Proceedings of the IEEE international conference on computer vision, 2015: 1520-1528.

[30] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 2881-2890.

[31] Zhao H, Qi X, Shen X, et al. Icnet for real-time semantic segmentation on high-resolutionimages [C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 405-420.

[32] Chen L-C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.

[33] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks [C]. Proceedings of the IEEE international conference on computer vision, 2017: 764-773.

[34] Ghiasi G, Fowlkes C C. Laplacian pyramid reconstruction and refinement for semantic segmentation [C]. European Conference on Computer Vision, 2016: 519-534.

[35] Lin G, Milan A, Shen C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 1925-1934.

[36] Li X, Liu Z, Luo P, et al. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 3193-3202.

[37] Pinheiro P H, Collobert R. Recurrent convolutional neural networks for scene labeling [C]. 31st International Conference on Machine Learning (ICML), 2014: 925-934.

[38] Shuai B, Zuo Z, Wang B, et al. Dag-recurrent neural networks for scene labeling [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 3620-3629.