CSR: A Lightweight Crowdsourced Road Structure Reconstruction System for Autonomous Driving

Huayou Wang*, Qingyao Liu*, Jiazheng Wu, Kun Liu, Chao Ding, Xianpeng Lang, Changliang Xue

***Abstract*— Highly accurate and robust vectorized reconstruction of road structures is crucial for autonomous vehicles. Traditional LiDAR-based methods require multiple processes and are often expensive, time-consuming, labor-intensive, and cumbersome. In this paper, we propose a lightweight crowdsourced road structure reconstruction system (termed CSR) that relies solely on online perceived semantic elements. Ambiguities and perceptual errors of semantic features and Global Navigation Satellite System (GNSS) global pose errors constitute the predominant challenge in achieving alignment across multi-trip data. To this end, a robust two-phased coarse-to-fine multi-trip alignment method is performed considering local geometric consistency, global topology consistency, intra-trip temporal consistency, and inter-trip consistency. Further, we introduce an incremental pose graph optimization framework with adaptive weight tuning ability to integrate pre-built road structures, currently perceived multi-trip semantic features, odometry, and GNSS, enabling accurate and robust incremental road structure reconstruction. CSR is highly automated, efficient, and scalable for large-scale autonomous driving scenarios, significantly expediting road structure production. We quantitatively and qualitatively validate the reconstruction performance of CSR in real-world scenes. CSR achieves centimeter-level accuracy commensurate with established LiDAR-based methods, concurrently boosting efficiency and reducing resource expenditure.**

I. INTRODUCTION

Autonomous driving has attracted widespread attention with the increasing adoption of Advanced Driver Assistance Systems. A current trend is to abandon high-definition (HD) maps in favor of real-time online Bird's Eye View (BEV) perception. However, this presents challenges at complex intersections, necessitating the reconstruction of road structures to provide prior information. Additionally, the reconstructed road structures can provide training labels for perception networks.

The recognition of the inefficiency and resource-intensive nature of traditional LiDAR-based road structure reconstruction methods [1], [2], especially their dependence on extensive manual annotation, has sparked significant interest in automated road structure generation. Rapid, accurate, and cost-effective road structure reconstruction is becoming increasingly critical to the industry, leading to a surge in research aimed at the automatic reconstruction of road structures.

As an increasing number of vehicles become outfitted with onboard sensors, including cameras and LiDAR, crowdsourcing emerges as a rapid and scalable method for gathering road structure data. The crowdsourced acquisition and reconstruction strategy boasts minimal computational load and rapid data transfer, while its data compactness ensures efficient storage utilization without compromising details necessary for practical application. Although some researchers [3] [4] have researched the direction of crowdsourced road structure reconstruction, it is still a challenge to build an algorithm that can align multi-trip data to generate a complete lane topology in various scenarios. Therefore, we propose a lightweight crowdsourced road structure reconstruction system that relies solely on online perceived road semantic elements, as shown in Fig. 1. The main contribution of this paper is summarized as follows:

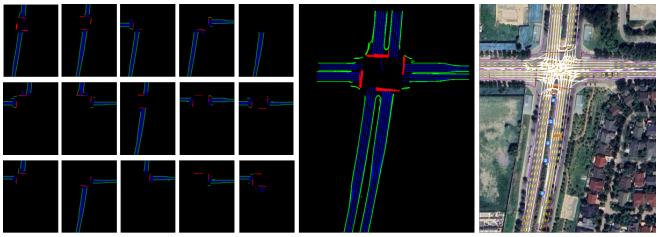


Fig. 1: Picture (a) shows the perceived semantic elements of each trip. Picture (b) displays the multi-trip alignment results. Picture (c) presents the instance lane topology generated by the proposed system automatically, and the road structure is aligned with Google Maps for visualization.

- A lightweight crowdsourced road structure reconstruction system that relies on vectorized semantic elements.
- A robust two-stage coarse-to-fine multi-trip alignment algorithm considering local geometric consistency, global topology consistency, intra-trip temporal consistency, and inter-trip consistency.
- An incremental pose graph optimization framework with adaptive weight tuning ability that integrates pre-built road structures, currently perceived multi-trip semantic features, odometry, and GNSS, enabling accurate and smooth incremental road structure reconstruction.
- Experiments on real-world autonomous driving scenes show that CSR achieves centimeter-level accuracy comparable to traditional LiDAR-based methods, but is more time-efficient and reduces the consumption of transmission, computational, and storage resources.

II. RELATED WORK

A. Online Road Structure Reasoning

Online road structure reasoning methods aim to infer the road structure within a certain range around the vehicle

*Equal Contribution.

All authors are with Li Auto Inc., Beijing, China. {wanghuayou, liuqingyao, wujiazheng, liukun, dingchao5, xuechangliang}@lixiang.com.

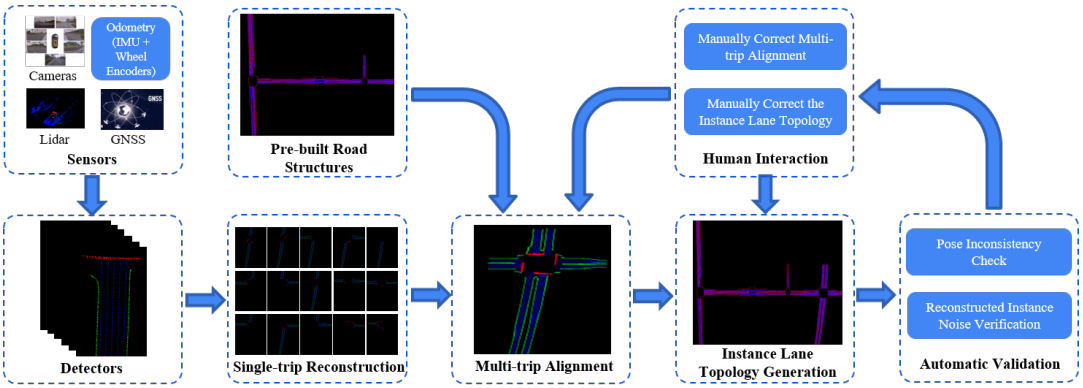


Fig. 2: Overview of the road structure automatic reconstruction framework. It comprises seven parts: sensors, detectors, single-trip reconstruction, multi-trip alignment, instance lane topology generation, automatic validation, and human interaction.

in real time. With the development of PV-to-BEV transformation [5]–[8] , HDMapNet [9] employs such transformation and a segmentation head to predict semantic features around the vehicle and extracts vector topology through post-processing. Subsequent models like VectorMapNet [10], Persformer [11], Pivotnet [12], Streammapnet [13] and MapTR [14] [15] reduce this post-processing, achieving end-to-end output of vectorized road elements. However, these methods can only obtain local-scale lane topology and struggle with complex scenes. In contrast, our work aims at creating a large-scale global consistent road structure topology and focuses on reconstruction quality.

B. LiDAR-based Road Structure Reconstruction Methods

Traditional LiDAR-based techniques [1], [2], [16]–[18] entail multipass point cloud registration and global pose optimization, wherein road elements are manually delineated from the LiDAR reflectivity data. To reduce time and labor costs, VMA [19] implements map annotation using the branch-and-bound method, and uses MapTR [14] to extract road elements from Lidar reflectivity data. Some methodologies [20] integrate image semantic segmentation with point clouds, followed by post-processing to derive 3D semantic point clouds and instance semantics. While these approaches facilitate high-accuracy semantic reconstruction, LiDAR-based algorithms demand substantial transmission, storage, and computational resources.

C. Crowdsourced Road Structure Reconstruction Methods

In the industrial domain, Mobileye has pioneered the deployment of 2D crowdsourcing technologies with its Road Experience Management system. Following this initiative, several other enterprises have shifted their focus towards similar advancements. Notable examples include HERE's Live Map, Bosch's Road Signature, and Horizon's NavNet.

In academia, Regder et al. [21] detected lanes from images and produced local grid maps through odometry, enhancing accuracy with map stitching techniques for pose optimization. Qiao et al. [22] introduced a lane reconstruction approach leveraging 3D lane perception and odometry to generate splines by solving lane association as an assignment problem with a bipartite graph. Qin et al. [23] built the semantic map of underground parking lots by road markers, and semantic point cloud stitching and loop closure were performed to build a globally consistent map. The previously stated methods can only process single-trip data with loop closure, making it difficult to construct a complete road topological structure.

Wijaya et al. [24] implemented crowdsourced point-based visual SLAM to align and combine the local maps derived by multiple vehicles. Herb et al. [25] also proposed a crowdsourcing approach to generate semantic maps. However, they are difficult to apply since inter-session image feature matching consumes a lot of computation and the feature points are not robust enough. Qin et al. [3] proposed a lightweight and highly feasible semantic reconstruction framework (RoadMap) in a crowdsourced way for visual localization. The semantic road information represented by semantic point clouds from multiple vehicles was fused directly based on GNSS. In order to solve the map blur problem of the RoadMap, in [4], a global point cloud registration network based on GeoTransformer [26] was introduced to optimize the results of RoadMap.

The previously introduced methods use semantic point clouds instead of vectorized representations, still require a large amount of transmission and storage resources, and require complex post-processing to generate semantic feature instances. In this paper, we use the BEV perceived semantic features with instance representation to perform incremental crowdsourced road structure reconstruction. In addition, none of the above methods carefully mentioned the semantic data association problem among multi-trips. Ambiguities and perceptual errors in semantic features and substantial initial GNSS inaccuracies present significant challenges in data association. To address this issue, a robust data association approach was employed, emphasizing local geometric consistency, global topology congruency, intra-trip temporal consistency, and inter-trip coherency.

III. METHODOLOGY

A. System Overview

The crowdsourced road structure reconstruction problem can be defined as: Given the perceived road semantic elements $\mathcal{Z} = \left\{\mathcal{Z}_k = \{\mathbf{z}_k^t\}_{t=1}^{T}\right\}_{k=1}^{K}$, GNSS measurements

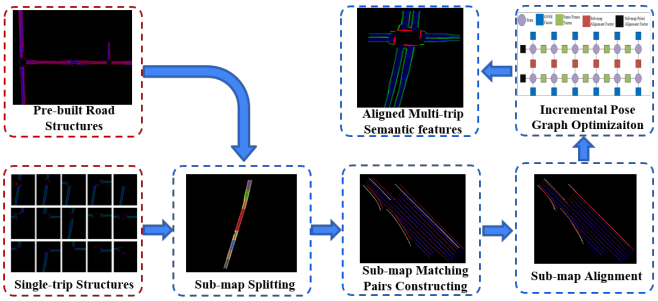


Fig. 3: Overview of the coarse alignment process.

$\mathcal{G} = \left\{\mathcal{G}_k = \{g_k^t\}_{t=1}^T\right\}_{k=1}^K$ and odometry measurements $\mathcal{O} = \left\{\mathcal{O}_k = \{o_k^t\}_{t=1}^T\right\}_{k=1}^K$ of multi-trip data, estimate the global pose $\mathcal{X} = \left\{\mathcal{X}_k = \{x_k^t\}_{t=1}^T\right\}_{k=1}^K$ and construct a unified vector map $\mathcal{L} = \{\mathbf{l}_m\}_{m=1}^M$. The pose x_k^t and landmark position $\mathbf{l}_m$ are defined as $x_k^t \in SE(2)$ and $\mathbf{l}_m \in \mathbb{R}^2$. The road structure reconstruction problem can be represented as the following maximum a posteriori (MAP) inference problem:

$$\hat{\mathcal{X}}, \hat{\mathcal{L}} = \arg\max_{\mathcal{X}} p(\mathcal{X}, \mathcal{L} | \mathcal{Z}, \mathcal{G}, \mathcal{O}). \tag{1}$$

Therefore, the CSR system is divided into seven components, namely sensors, detectors, single-trip reconstruction, multi-trip alignment, instance lane topology generation, automatic validation, and human interaction as shown in Fig. 2. Sensors consist of cameras, LiDAR, GNSS, and odometry (IMU and wheel encoders). The cameras and LiDAR are used to detect semantic elements. The odometry provides local relative motion estimation. The detector layer detects road center lines, lane dividers, road boundaries, stop lines, road markings, crosswalks, etc. The single-trip reconstruction layer performs tracking, smoothing, and splicing of semantic features perceived in consecutive frames. The multi-trip alignment layer aligns multi-trip perceived semantic elements into one global coordinate. In the instance lane topology generation layer, the aligned multi-trip perceived semantic features are merged to generate the instance lane topology. The automatic validation layer evaluates the quality of both the multi-trip alignment and instance lane topology generation layer. The human interaction layer addresses instances of algorithmic failure by manually refining the outcomes of multi-trip alignment and instance lane topology generation.

B. Semantic Features and Detection

The semantic elements that need to be extracted can be divided into three categories according to the representation form: line-type elements, bounding box-type elements, and closed-type elements. In this paper, we adopt the MapTR [14], [15] algorithm used in VMA [19] to extract semantic elements, because the method can model various road semantic elements in a unified point sequences manner.

C. Single-Trip Semantic Reconstruction

The single-trip semantic reconstruction procedure mainly comprises semantic element tracking, denoising, splicing, and pose optimization. This process is independently and parallelly conducted inside each trip. For the perceived semantic features, we employ the state-of-the-art (SOTA) Catmull-Rom Spline-based lane reconstruction method proposed by Qiao [22]. This method utilizes a bipartite graph to model the lane association process as an assignment problem and achieves edge weight assignment by incorporating Chamfer Distance (CD), attitude uncertainty, and lateral sequence consistency. Additionally, this algorithm carefully designs control point initialization, spline parameterization, and optimization to create, extend, and refine splines step by step. Through these careful designs, this step can denoise the perception results of consecutive frames and obtain smooth single-trip splicing outcomes. In this step, we will also extract keyframes based on movement distance and period for subsequent steps to reduce calculations.

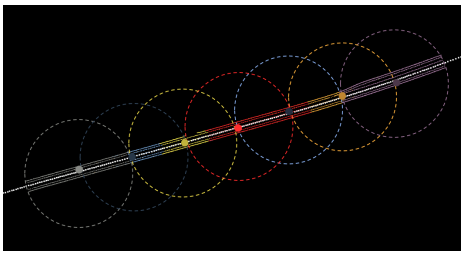


Fig. 4: The sub-map splitting process. The single-trip data within a circular area forms a sub-map, and the sub-maps in the same circle of different trips constitute matching pairs.

Single-trip Pose Smooth: To solve the GNSS pose jumping problem within a single trip, particularly in challenging environments such as tunnels, elevated structures, and tree-lined streets, we integrate GNSS with odometry by constructing a pose graph optimization problem as follows:

$$\begin{aligned}\hat{\mathcal{X}}_k = \arg\min_{\mathcal{X}_k} \sum_t & ||e^o(x_k^t, x_k^{t+1}, o_k^{t,t+1})||^2_{\Omega_k^{o,t}} \\ & + ||e^g(x_k^t, g_k^t)||^2_{\Omega_k^{g,t}}.\end{aligned} \tag{2}$$

The global pose x_k^t of frame t inside trip k is defined as a node of the graph. The relative pose $o_k^{t,t+1}$ of adjacent frames inside each trip is defined as the relative constraint edge, and the GNSS pose g_k^t of each frame is defined as the abstract constraint edge. Based on these edges, we can obtain optimized smooth poses and stitch the perception results based on the optimized poses of each trip.

D. Multi-Trip Alignment

Since the perceived features of each trip only contain a local area around the vehicle's trajectory, the fusion of multi-trip perception results is crucial to constructing the complete road structure of the autonomous driving scene. The main challenge in achieving multi-trip alignment is the existence of ambiguities and perceptual errors of semantic features and the GNSS pose error. Therefore we proposed a robust two-stage coarse-to-fine multi-trip alignment framework and an incremental pose graph optimization framework with adaptive weight tuning ability, enabling accurate and robust multi-trip alignment and incremental road structure reconstruction.

1) Coarse Alignment: The coarse alignment process is composed of four steps, including sub-map splitting, sub-map matching pair construction, sub-map alignment, and incremental global pose optimization, as shown in Fig. 3.

Algorithm 1 Sub-map Data Association Algorithm

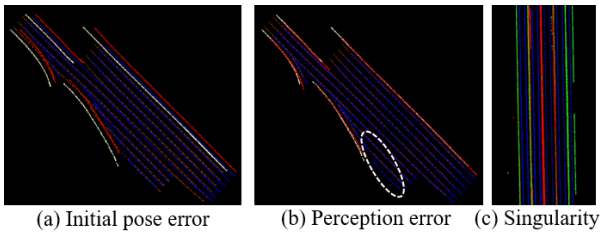
Input: $\mathcal{X}_{sub}^{0}$: initial pose of sub-maps; $\mathcal{M}_{sub}$: sub-maps;
Output: $\mathcal{D}^{*}$: optimal semantic feature data association

1: assess global topology consistency according to Eq. 3;
2: sliding window temporal consistency voting for the candidate sub-map matching pairs resulted from step 1.
3: construct multi-order graph matching problem, and compute $\mathcal{D}$ according to Eq. 4;
4: compute sub-map relative pose and matching confidence $c_{kk'}$ based on $\mathcal{D}$ according to Eq. 7;
5: perform intra-trip temporal smoothing according to Eq. 8 to gain temporal consistent $\mathcal{D}$;
6: performs inter-trip consistency check according to graph and maximum consistency set to obtain $\mathcal{D}^{*}$;
7: **return** $\mathcal{D}^{*}$

Sub-map Splitting and Sub-map Matching Pair Constructing. The process first constructs sub-maps by stitching the perceived features of consecutive frames to reduce the singularity of single-frame perception results. In order to balance the singularity and geometric accuracy of the sub-maps, we use the smoothed pose $\mathcal{G}_s = \left\{\mathcal{G}_{s,k} = \left\{g_{s,k}^{t}\right\}_{t=1}^{T}\right\}_{k=1}^{K}$ from single-trip reconstruction procedure to splice the sub-maps at a fixed radius distance. We construct sub-maps and sub-map matching pairs by building a graph, as shown in Fig. 4. The graph treats each frame as a node and constructs edges by pairing consecutive frames within a trip as well as frame pairs whose distance and angle between trips are less than a pre-defined threshold. Then a node in the graph that has not been divided into sub-maps is selected, and the frames within a fixed radius range from the node in the complete graph are used to construct a sub-map M_{sub} for each trip and the sub-maps of different trips in the same circle are treated as sub-map matching pairs $P_{sub-sub}$. The constructed matching pairs can ensure the overlap area is large enough to reduce the singularity of sub-map matching. If the segmented sub-map has sufficient overlap area with the pre-built road structures, then the sub-map and the overlap area of the pre-built road structures will be constructed as matching pairs $P_{sub-pre}$.

Sub-map Alignment. Because of the singularity and perception error of semantic features, and significant initial global pose error, data association (DA) becomes one of the most challenging problems for the sub-map alignment process, as shown in Fig. 5. In this paper, we propose a robust data association method based on local geometric, global topology, intra-trip temporal, and inter-trip consistency to solve the problems mentioned above of DA and ensure spatial and temporal consistency. To illustrate the proposed DA method, the pseudocode is provided in Algorithm 1. The details of the DA process are as follows:

Step 1: Preliminary Assessment of Global Topology. To accommodate the existence of various errors and singularities, an association based on global topological consistency is performed to determine whether two sub-maps $(M_{sub}^{i}, M_{sub}^{j})$ belong to the identical road area. Global road topological consistency refers to the global topologies (number of lanes $N_{nl}()$, road width $N_{rw}()$, split and merge points $N_{sm}()$, etc.) of the two sub-maps are similar and cannot have major deviations, and the distance of the two sub-maps road center N_{rc} is less than threshold τ_{rc}. The previous conditions can be formulated as:



(a) Initial pose error (b) Perception error (c) Singularity

Fig. 5: Difficulty in sub-map alignment scenarios.

$$\begin{cases} |N_{nl}(M_{sub}^{i}) - N_{nl}(M_{sub}^{j})| <= \tau_{nl}, \\ |N_{rw}(M_{sub}^{i}) - N_{rw}(M_{sub}^{j})| <= \tau_{rw}, \\ |N_{sm}(M_{sub}^{i}) - N_{sm}(M_{sub}^{j})| <= \tau_{sm}, \\ |N_{rc}(M_{sub}^{i}) - N_{rc}(M_{sub}^{j})| <= \tau_{rc}. \end{cases} \tag{3}$$

Since the semantic feature extraction algorithm has different accuracy in different scenarios, this step uses different parameter thresholds for different scenarios. These thresholds are established based on the noise estimation result of each element obtained through the single-trip reconstruction step. The greater the noise, the larger the parameter thresholds.

Finally, the binary outcomes $x_{kk'}$ obtained in this step for all candidate matching pairs between any two trips are voted by sliding window voting algorithm consistently in time sequence to eliminate some false candidate matching pairs.

Step 2: Matching based on Local Geometric Consistency. Local geometric consistency means that the relative geometric position relationships of the local adjacent elements are consistent. The candidate sub-map pair $P_{sub-sub}^{i,j}$ from step 1 is first aligned through the left and right road boundary matching to eliminate large GNSS initial pose errors. Then the lane center lines, stop line, road boundaries, and all other road elements are used to refine the sub-map alignment pose $\mathbf{T}_{sub-sub}^{i,j}$. Since the road boundary is generally far from the self-vehicle, there is a large error in perception, so the lateral pose error based on the aligned pose of the road boundary may still exceed half a lane, causing the singularity of lane line matching. Therefore, an optimal data association method considering the matching number, semantic feature similarity, and local geometric similarity is performed to achieve optimal global consistent matching. It is formulated as a multi-order graph matching problem:

$$\begin{aligned} \hat{\mathbf{X}} = \arg\max_{\mathbf{X}} \omega_1 N_m + \omega_2 \frac{1}{N_m} \sum_{(k,k')}^{N_m} x_{kk'} s_{kk'} \\ + \omega_3 \frac{1}{N_e} \sum_{(kk',ll')}^{N_e} x_{kk'} x_{ll'} s_{kl,k'l'}, \end{aligned} \tag{4}$$

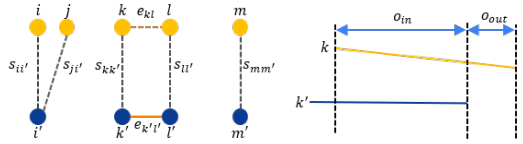


Fig. 6: Representation of element and edge similarity.

(a) (b)

Fig. 7: Figure (a) shows all possible correspondence pairs. Figure (b) presents the obtained maximum consistency set.

S.t.

$$\sum_{k=1}^{N_{sub}^{i}} x_{kk'} \leq 1, \ \sum_{k'=1}^{N_{sub}^{j}} x_{kk'} \leq 1, \ x_{kk'} = 0 \ or \ 1,$$
$$x_{kk'} = 0 \iff s_{kk'} < \tau_s,$$

where N_{sub}^{i} and N_{sub}^{j} are the number of perceived elements in sub-map M_{sub}^{i} and sub-map M_{sub}^{j}, N_m is the number of match pairs to be solved, N_e is the number of edges between the constructed matching pairs. ω_1, ω_2 and ω_3 are hyperparameters. $x_{kk'}$ denotes whether element k is matched with element k'. $s_{kk'}$ represents the similarity between element k and element k', it consists of position and overlap similarity:

$$\begin{aligned} s_{kk'} =& \omega \exp(-\frac{1}{2}(\frac{d_p^{kk'}}{\sigma_p})^2) \\ &+(1-\omega)\exp(-\frac{1}{2}(\frac{d_o^{kk'}}{\sigma_o})^2), \end{aligned} \tag{5}$$

where ω is a learned hyperparameter for weighting position similarity and overlap similarity. $d_p^{kk'}$ and $d_o^{kk'} = o_{in}/(o_{in} + o_{out})$ denote the Chamfer Distance and overlap ratio of two elements, as shown in Fig. 6. σ_p and σ_o can be learned from single-trip reconstruction results. $s_{kl,k'l'}$ represents the similarity between edge e_{kl} and $e_{k'l'}$:

$$s_{kl,k'l'} = \exp(-\frac{1}{2}(\frac{e_{kl} - e_{k'l'}}{\sigma_e})^2), \tag{6}$$

where e_{kl} and $e_{k'l'}$ denotes the distance between feature k and l, and feature k' and l' , as shown in Fig. 6 . σ_e can be learned offline.

The initial element matching results are obtained by solving Eq. 4 using the general random re-weighted walk framework [27]. Then the sub-map alignment pose can be solved by an iterative optimization algorithm based on the Graduated Non-Convexity (GNC) algorithm [28]. After the sub-maps are aligned, the confidence of the initial sub-map alignment is calculated based on the alignment pose transformation. The initial matching confidence $c_{kk'}$ is obtained through feature similarity and local geometric similarity:

$$c_{kk'} = \omega s_{kk'} + (1-\omega)\frac{1}{N_m}\sum_{ll'}^{N_m} s_{kl,k'l'}. \tag{7}$$

Step 3: Intra-trip Temporal Consistency. Intra-trip temporal consistency is that the matching results of the same elements of sub-maps in the same trip should be consistent. The matching correctness of an element in a sub-map can be verified by other sub-maps containing the same element in the same trip. Further, a mismatch in a sub-map can be found and corrected based on adjacent sub-map matching results. Intra-trip temporal smoothing acquires corresponding element k' of element k by weighting the candidate matching elements $D_{1:N}$ and matching confidence $c_{k,n}$ over each sub-map inside the trip.

$$\hat{k'} = \arg\max_{k'} \frac{x_{kn}c_{k,n}}{\sum_n x_{kn}c_{k,n}}. \tag{8}$$

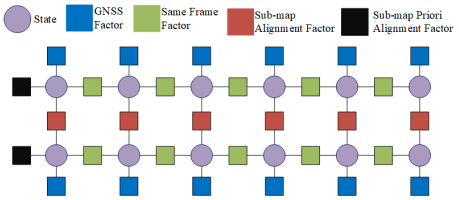


Fig. 8: Factor graph of multi-trip pose graph optimization.

Step 4: Inter-trip Consistency. After obtaining the matching results of the sub-maps between two trips, we perform a graph-based method to perform inter-trip consistency verification on the matching results of the sub-maps between all trips, as shown in Fig . 7. The graph treats each matching pair as a node and constructs edges with matching pairs that have at least one identical node to that matching pair. Then the maximum consistent set of each feature in the graph is obtained for outlier pruning by finding the maximum clique of the graph using a fast exact parallel maximum clique extraction algorithm [29]. After getting the matching inlier, we utilize the GNC algorithm again to solve the relative pose of the sub-map matching pair and recalculate the matching confidence through Eq. 7.

Incremental Pose Graph Optimization. After obtaining the sub-map alignment relative poses, an incremental pose graph optimization algorithm with adaptive weight tuning is executed to obtain the aligned global poses of all N_{sub} sub-maps. We formulate a non-linear least square estimator based on odometry measurement, smoothed global initial pose g_{sub}^{i} of sub-map M_{sub}^{i}, sub-map pair $(M_{sub}^{i}, M_{sub}^{j})$ registration relative pose $T_{reg}^{i,j}$ and global pose T_{pre}^{i} obtained by aligning the sub-map M_{sub}^{i} to pre-built road structures to estimate the pose of each sub-map $\mathcal{X}$. The optimization objective is represented as follows:

$$\begin{aligned} \hat{\mathcal{X}}_{sub} = \arg\min_{\mathcal{X}_{sub}} & \sum_{i}^{N_{sub}} ||e^{g}(x_{sub}^{i}, g_{sub}^{i})||_{\Omega_g^i}^2 \\ &+ \sum_{(i,pre)}^{P_{sub-pre}} ||e^{pre}(x_{sub}^{i}, T_{pre}^{i})||_{\Omega_{pre}^i}^2 \\ &+ \sum_{(i,j)}^{P_{sub-sub}} ||e^{reg}(x_{sub}^{i}, x_{sub}^{j}, T_{reg}^{i,j})||_{\Omega_{reg}^{i,j}}^2 \\ &+ \sum_{i}^{N_{sub}} \sum_{t}^{M_{sub}^{i}} \sum_{j,j\neq i, t\in M_{sub}^{j}}^{N_{sub}} ||e^{sm}(x_{sub}^{i}, x_{sub}^{j}, T_i^t, T_j^t)||_{\Omega_{sm}}^2, \end{aligned} \tag{9}$$

where each error term together with the corresponding information matrix can be regarded as a factor, and each state variable can be regarded as a node, therefore the incremental

pose graph optimization problem can be expressed by factor graph, as shown in Fig. 8. Error terms are composed of global absolute pose error e^g, global pose constraints e^{pre} aligned to the prior map, sub-map alignment relative pose error e_{reg}, and consistent constraints on global poses in different sub-maps for the same frame e^{sm}. The four constraints are defined as:

$$\begin{cases} e^g(x^i_{sub}, g^i_{sub}) = {x^i_{sub}}^{\mathrm{T}} g^i_{sub} \\ e^{pre}(x^i_{sub}, T^i_{pre}) = ({x^i_{sub}}^{\mathrm{T}} x^p_i) T^i_{pre} \\ e^{reg}(x^i_{sub}, x^j_{sub}, T^{i,j}_{reg}) = ({x^i_{sub}}^{\mathrm{T}} x^j_{sub}) T^{i,j}_{reg} \\ e^{sm}(x^i_{sub}, x^j_{sub}, T^t_i, T^t_j) = (x^i_{sub} T^t_i)^{\mathrm{T}} (x^j_{sub} T^t_j), \end{cases} \tag{10}$$

where T^t_i and T^t_j represent the pose of frame t in sub-map M^i_{sub} and sub-map M^j_{sub} respectively. The non-linear optimization problem is solved directly by Gauss-Newton iterative algorithm.

Adaptive Weight Tuning: The residual weight Ω^i_g of the global initial pose g^i_{sub} is set based on the status of GNSS s^t_i and deviation of relative positions d^t_{g-o} of GNSS and odometry in consecutive frames. It can be represented as:

$$\Omega^i_g = \frac{1}{N_t} \sum_t^{N_t} \left\{ M_{table}(s^t_i) + \omega d^t_{g-o} \right\} \Omega^{base}_g, \tag{11}$$

where M_{table} is a table that stores the corresponding relationship between GNSS status and weight. This table is obtained through statistics of real data. Ω^{base}_g is the base information matrix of the global initial pose residual. The residual weight of e_{sm} is set based on the statistical accuracy of the odometry in various scenarios. The weights of e_{pre} and e_{reg} are set based on the sub-map matching confidence c_{ij} calculated in the previous step:

$$\Omega^{i,j}_{reg} = \alpha c_{ij} \Omega^{base}_{reg}. \tag{12}$$

2) Fine Alignment: Since the sub-map itself still has certain errors, there are still errors in the pose graph optimization results based on the sub-map level, so precise alignment at the frame level is necessary. This fine alignment step uses the same algorithm flow as the previous coarse alignment step, except that the processing primitive is replaced by single frame perception results from a sub-map. Through this step, accurate global optimized poses are obtained.

E. Lane Instance Topology Generation

The multi-trip perception results are spliced together based on the globally aligned optimized pose obtained in the previous step. The semantic elements of different trips are matched together, and the multi-pass perception results are fused into instances using the same method as the single-trip semantic reconstruction procedure. Then the obtained instance geometry and variance of noise are used to determine whether there is noise in the map. If there are noises in the map, we will use human interaction tools similar to interactive SLAM [30] to continuously repair wrong matching pairs, add missing matching pairs manually and give a larger weight, and finally obtain the corrected pose by resolving the optimization problem of Eq. 9. Then the lane instance topology generation process is re-executed based on the corrected poses.

Fig. 9: Overview of experimental data from a satellite view.

IV. Experiment

Quantitative and qualitative assessments were performed in real-world scenarios to evaluate the proposed CSR system. The empirical findings demonstrate that our approach achieves centimeter-level accuracy and requires less time and fewer storage resources compared to traditional LiDAR-based methods.

A. Implementation Details

1) Data Source: We carried out experiments utilizing real-world, crowdsourced data derived from autonomous driving test areas. Fig. 9 shows several scenes in our dataset.

2) Experimental Setup: The ground truth (GT) is derived from the multi-sensor fusion reconstruction method. Firstly, the multi-sensor data including LiDAR, Cameras, GNSS, and Odometry collected from vehicles is fed into a multi-sensor fusion reconstruction pipeline to obtain pose estimation results. The instance lane topology generation method presented in this paper is applied to derive the ultimate lane topological structure. Moreover, to validate result accuracy and establish ground truth suitability, the reconstructed road structures underwent meticulous manual correction to reinforce their precision.

3) Evaluation Metrics: The evaluation of the CSR system employs a dual-faceted strategy, focusing on both the precision of pose optimization and the quality of the final reconstruction outcomes. Pose accuracy is assessed through Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) computations, with deviations measured via Root Mean Square Error (RMSE) and Mean Error (ME). These metrics are averaged over the trajectory's extent to encapsulate overall system precision. We employ the CD to assess the quality of the reconstructed road structures.

B. Quantitative results

1) Ablation Study: To demonstrate the nuanced efficacy of the proposed coarse-to-fine multi-trip alignment method within the CSR system, we compare three kinds of designs, coarse alignment, fine alignment, and our integrated coarse-to-fine alignment approach. As shown in Table. I, it depicts that the coarse-to-fine alignment approach achieves higher pose estimation accuracy. Coarse alignment demonstrated superior pose estimation results compared to fine alignment, primarily attributed to the latter's reliance on single-frame

TABLE I: **Ablation study of coarse-to-fine alignment**.

Coarse Alignment	Fine Alignment	RPE [m]↓	
		RMSE	ME
✓		0.096	0.064
	✓	0.111	0.078
✓	✓	**0.087**	**0.056**

* The best results are highlighted in bold.

TABLE II: **Comparison of lane topology reconstruction.**

Method	CD_{ld}	CD_{cl}	CD_{bdr}	mCD
RoadMap++ [4]	0.0636	0.0611	0.0824	0.0690
CSR **(Ours)**	**0.0521**	**0.0537**	**0.0681**	**0.0580**

* ld denotes lane dividers, cl represents center lines, and bdr means road boundaries. The unit of measurement is meters (m).

outcomes, which struggles with submap alignment singularities, significant perception errors, and notable GNSS pose inaccuracies. The errors in coarse alignment predominantly stem from inaccuracies inherent in sub-map stitching.

TABLE III: **Comparison of pose optimization.**

Method	ATE [m]↓		RPE [m]↓	
	RMSE	ME	RMSE	ME
RoadMap++ [4]	0.194	0.177	0.107	0.076
CSR **(Ours)**	**0.155**	**0.135**	**0.087**	**0.056**

2) Comparison with the SOTA Method: To further illustrate the superiority of the CSR system, we benchmark our method against a traditional LiDAR-based method and the SOTA crowdsourced road structure reconstruction framework RoadMap++ [4] which is an improved version of the original RoadMap [3]. Due to its proprietary nature, we reproduced its method from the original description. As shown in Table. III, our CSR achieves higher performances on all pose evaluation metrics. The RPE metric is lower than ATE, RPE is prioritized as it reflects the relative accuracy vital for road structure reconstruction. And in Table.II, CSR achieves higher instance lane topology reconstruction accuracy for center lines, lane dividers, and road boundaries. Enhanced performance is attributable to adopting a multi-stage coarse-to-fine alignment process and a sophisticated sub-map matching algorithm that integrates global topological structure, local geometric relationship, and intra-trip chronological and inter-trip coherence. Conversely, Roadmap++ utilizes GeoTransfomer [26] for aligning point clouds, an approach that neglects instance-specific features and presents challenges in enforcing both intra-trip temporal and inter-trip validation.

C. Qualitative Results

1) Coarse-to-Fine Alignment: To visualize the capability of our coarse-to-fine alignment framework, we present a typical complicated scene with initial pose error and road singularity, as shown in Fig.10. It can be seen that our coarse-to-fine alignment method achieves the best alignment performance. Coarse alignment can roughly align the multi-trip trajectories, but there are still errors in the alignment results. Fine alignment focuses more on frame-to-frame alignment with limited alignment ability under bad initial

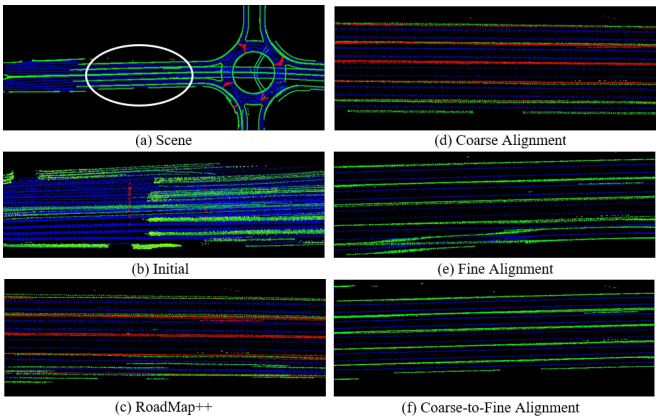


Fig. 10: Multi-trip alignment results of different experimental setup. Zoom in for a better view.

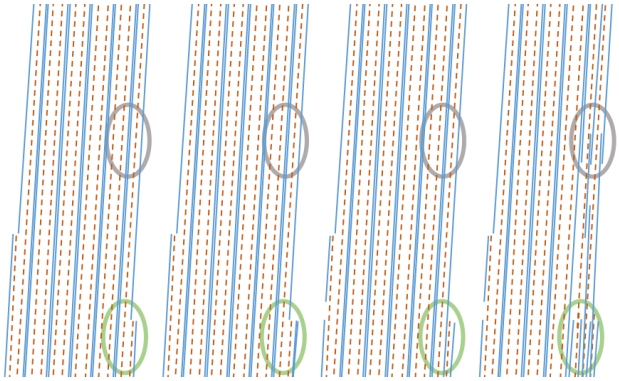


Fig. 11: Instance lane topology generation results.

pose and singularity cases. RoadMap++ matches different lanes due to its inability to solve the singularity of elements.

2) Instance Lane Topology Generation: To visualize the final results of instance lane topology generation, we compared CSR with the traditional LiDAR-based method and RoadMap++. All these methods use the same instance lane topology generation algorithm as our proposed system, with the sole difference residing in the multi-trip pose optimization process. As demonstrated in Fig.11, our CSR achieves highly accurate results, similar to the traditional LiDAR-based method, while RoadMap++ generates chaotic instance lane topology. This primarily benefits from the coarser-to-fine multi-trip alignment framework, which is equipped with a robust sub-map semantic feature data association algorithm.

D. Source Cost

We conducted a quantitative analysis of all collected data storage requirements and processing time. The detailed results are presented in TABLE IV. It can be observed that CSR significantly reduces memory and time consumption compared with the traditional LiDAR-based method.

V. CONCLUSIONS

We present CSR, a lightweight crowdsourced road structure reconstruction framework based on the online perceived semantic features, the coarse-to-fine multi-trip alignment method, and the incremental pose graph framework with

TABLE IV: **Comparison of source cost.**

Cost	Traditional method	CSR
Storage (Gb/km)	2.4	**0.027**
Time (min/km)	5.881	**0.285**

adaptive weight tuning. CSR is highly automated, efficient, and scalable for large-scale autonomous driving scenarios, significantly expediting road structure reconstruction and curbing costs with minimal storage and computing resources. CSR achieves centimeter-level accuracy comparable to traditional LiDAR-based methods, significantly improving efficiency and reducing resource costs. This greatly proves the value of CSR in providing navigation priors and perceptual annotation data. We find that the quality of semantic feature extraction greatly affects the final results, therefore we will first optimize our semantic feature extraction algorithm in the future. The instance lane topology generation algorithm will also be optimized using SOTA large-scale lane graph construction algorithms, such as LaneGAP [31], DAGMapper [32], etc.

REFERENCES

[1] S. Yang, X. Zhu, X. Nian, L. Feng, X. Qu, and T. Ma, "A robust pose graph approach for city-scale lidar mapping," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1175–1182.

[2] Y. Wang, Y. Lou, W. Song, and Z. Tu, "A tightly-coupled framework for large-scale map construction with multiple non-repetitive scanning lidars," *IEEE Sensors Journal*, vol. 22, no. 4, pp. 3626–3636, 2022.

[3] T. Qin, Y. Zheng, T. Chen, Y. Chen, and Q. Su, "A light-weight semantic map for visual localization towards autonomous driving," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 248–11 254.

[4] T. Qin, H. Huang, Z. Wang, T. Chen, and W. Ding, "Traffic flow-based crowdsourced mapping in complex urban scenario," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 5077–5083, 2023.

[5] Y. Ma, T. Wang, X. Bai, H. Yang, Y. Hou, Y. Wang, Y. Qiao, R. Yang, D. Manocha, and X. Zhu, "Vision-centric bev perception: A survey," *arXiv preprint arXiv:2208.02797*, 2022.

[6] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 194–210.

[7] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*. Springer, 2022, pp. 1–18.

[8] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

[9] Q. Li, Y. Wang, Y. Wang, and H. Zhao, "Hdmapnet: An online hd map construction and evaluation framework," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 4628–4634.

[10] Y. Liu, T. Yuan, Y. Wang, Y. Wang, and H. Zhao, "Vectormapnet: End-to-end vectorized hd map learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 22 352–22 369.

[11] L. Chen, C. Sima, Y. Li, Z. Zheng, J. Xu, X. Geng, H. Li, C. He, J. Shi, Y. Qiao, *et al.*, "Persformer: 3d lane detection via perspective transformer and the openlane benchmark," in *European Conference on Computer Vision*. Springer, 2022, pp. 550–567.

[12] W. Ding, L. Qiao, X. Qiu, and C. Zhang, "Pivotnet: Vectorized pivot learning for end-to-end hd map construction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3672–3682.

[13] T. Yuan, Y. Liu, Y. Wang, Y. Wang, and H. Zhao, "Streammapnet: Streaming mapping network for vectorized online hd map construction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7356–7365.

[14] B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, W. Liu, and C. Huang, "Maptr: Structured modeling and learning for online vectorized hd map construction," *arXiv preprint arXiv:2208.14437*, 2022.

[15] B. Liao, S. Chen, Y. Zhang, B. Jiang, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Maptrv2: An end-to-end framework for online vectorized hd map construction," *arXiv preprint arXiv:2308.05736*, 2023.

[16] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time." in *Robotics: Science and systems*, vol. 2, no. 9. Berkeley, CA, 2014, pp. 1–9.

[17] T. Shan and B. Englot, "Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4758–4765.

[18] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and D. Rus, "Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping," in *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2020, pp. 5135–5142.

[19] S. Chen, Y. Zhang, B. Liao, J. Xie, T. Cheng, W. Sui, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Vma: Divide-and-conquer vectorized map annotation system for large-scale driving scene," *arXiv preprint arXiv:2304.09807*, 2023.

[20] K. Tang, X. Cao, Z. Cao, T. Zhou, E. Li, A. Liu, S. Zou, C. Liu, S. Mei, E. Sizikova, *et al.*, "Thma: Tencent hd map ai system for creating hd map annotations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 13, 2023, pp. 15 585–15 593.

[21] E. Rehder and A. Albrecht, "Submap-based slam for road markings," in *2015 IEEE Intelligent Vehicles Symposium (IV)*, 2015, pp. 1393–1398.

[22] Z. Qiao, Z. Yu, H. Yin, and S. Shen, "Online monocular lane mapping using catmull-rom spline," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7179–7186.

[23] T. Qin, T. Chen, Y. Chen, and Q. Su, "Avp-slam: Semantic visual mapping and localization for autonomous vehicles in the parking lot," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5939–5945.

[24] B. Wijaya, K. Jiang, M. Yang, T. Wen, X. Tang, and D. Yang, "Crowdsourced road semantics mapping based on pixel-wise confidence level," *Automotive Innovation*, vol. 5, no. 1, pp. 43–56, 2022.

[25] M. Herb, T. Weiherer, N. Navab, and F. Tombari, "Crowd-sourced semantic edge mapping for autonomous vehicles," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 7047–7053.

[26] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, and K. Xu, "Geometric transformer for fast and robust point cloud registration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 143–11 152.

[27] J. Lee, M. Cho, and K. M. Lee, "Hyper-graph matching via reweighted random walks," in *CVPR 2011*. IEEE, 2011, pp. 1633–1640.

[28] H. Yang, P. Antonante, V. Tzoumas, and L. Carlone, "Graduated non-convexity for robust spatial perception: From non-minimal solvers to global outlier rejection," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1127–1134, 2020.

[29] R. A. Rossi, D. F. Gleich, and A. H. Gebremedhin, "Parallel maximum clique algorithms with applications to network analysis," *SIAM Journal on Scientific Computing*, vol. 37, no. 5, pp. C589–C616, 2015.

[30] K. Koide, J. Miura, M. Yokozuka, S. Oishi, and A. Banno, "Interactive 3d graph slam for map correction," *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 40–47, 2020.

[31] B. Liao, S. Chen, B. Jiang, T. Cheng, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Lane graph as path: Continuity-preserving path-wise modeling for online lane graph construction," *arXiv preprint arXiv:2303.08815*, 2023.

[32] N. Homayounfar, W.-C. Ma, J. Liang, X. Wu, J. Fan, and R. Urtasun, "Dagmapper: Learning to map by discovering lane topology," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2911–2920.