Multi-model Traffic Scene Simulation with Road Image Sequences and GIS Information

Zhichao Cui*, Yuehu Liu§, Fuji Ren†, Qilin Zhang‡

***Abstract*— In this paper, a new multi-modal traffic scene simulation framework with combined inputs of road image sequences and road information from Geographic Information Systems (GIS) is proposed. The proposed framework contains two major steps, with the first one being a preprocessing step, including 3D road model extraction, camera location and orientation estimation and lane extraction from both GIS and road image sequences. After such preprocessing, the traffic scene reconstruction is reformulated into a 6-degree of freedom (6DoF) pose estimation in the 3D road model. Subsequently, the iterative closest point (ICP) algorithm is exploited for coarse point registration by estimating the pose in the road model. In addition, an objective function is established to incorporate the image features (e.g., lanes) into the road model and to refine the pose estimation. In the experiments with the publicly available KITTI dataset, the proposed method achieves high average Intersection-over-Union (IoU) scores as compared to the ground truth image sequences.**

I. INTRODUCTION

In recent years, the autonomous vehicle technology has evolved rapidly, thanks to the prevalence of deep learning-based algorithms [1]–[4]. Therefore, cost-efficient and repeatable evaluation methods of such autonomous driving systems is necessary to promote robustness and safety. Currently, one of the most popular forms of such evaluations is field test. Especially, institutions, corporations, colleges and government agencies have cooperated to organize multiple unmanned vehicle field contests, such as Defense Advanced Research Projects Agency (DARPA) Grand Challenge, AI and Intelligent Vehicles Future Challenge (IVFC), to promote the development of autonomous driving technology. Granted that such contests are valuable, field tests are generally prohibitively expensive, time-consuming and generally non-repeatable (e.g., due to mutable weather/traffic conditions). Therefore, it is necessary to test autonomous driving systems in a controlled, simulated environment. In addition, simulation based evaluation framework can readily create rare and potentially dangerous traffic scenes, such as in severe accidents and inclement weather (e.g., hurricane/blizzard), to verify the autonomous vehicles ability of handling emergency.

*Z. Cui is with Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an City, 710049, China; also with School of Advanced Technology and Science, Tokushima University, Tokushima, 770-8501, Japan. (e-mail: cui.zhichao@stu.xjtu.edu.cn)

§Y. Liu is with Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an City, 710049, China. (e-mail: liuyh@xjtu.edu.cn)

†F. Ren is with School of Advanced Technology and Science, Tokushima University, Tokushima, 770-8501, Japan. (e-mail: ren@is.tokushima-u.ac.jp)

‡Q. Zhang is with HERE Technologies, 425 W Randolph St, Chicago, Illinois, 60606, USA. (e-mail: samqzhang@gmail.com)

Z. Cui is the corresponding author.

In general, real traffic scenes consist of both static elements (e.g., roads, traffic signs, traffic lights) and dynamic elements (e.g., vehicle, pedestrians). The objective of traffic simulation is to reconstruct all the traffic elements in space-time continuum. Based on the technologies and source data, the virtual traffic scene reconstruction methods can be categorized into two categories [5]. The first category of methods are based on computer graphics and virtual reality technologies, and rely heavily on 3D graphics engine. For example, the PreScan software developed by TASS International Corporation can integrate traffic elements such as roads, traffic signs, vehicles and pedestrians, to construct various traffic scenes. However, such fully synthetic virtual traffic scenes are commonly oversimplified as compared to real traffic scenes, therefore, simulation tests based on oversimplified scenes might be unconvincing in evaluating autonomous driving systems.

Alternatively, another category of traffic scene reconstruction methods are based on multi-sensors data of real traffic scenes, which are either captured by vehicle equipped with multi-sensors (KIT vehicle [6]) or geo-sensing satellites. Nico Cornelis et al. [7] reconstructed 3D urban traffic scenes containing road, buildings and objects with two video streams and Global Positioning System (GPS) information. Li et al. [8], [9] reconstructed 3D traffic scenes from traffic image sequences. The simulated virtual traffic scene in [8], [9] has implemented functionalities such as virtual touring at various speed, virtual touring in any perspective (i.e., from arbitrary camera location and orientation) and panorama views.

In this paper, we utilize multi-model data (road image sequence and GIS information) to construct traffic scenes, especially the road ways, where the location, orientation and physical dimensions of these simulated roadways accurately accord with those from the road image sequences after reprojection. Unlike traditional 3D reconstruction methods, the proposed new framework reconstructs simulated traffic scenes through estimating the location and orientation of the roadway (i.e., 6DOF estimation), including the camera location/orientation and the road model. Through refining the pose of road model to accord with the image sequence, accurate registration can be achieved in the simulated traffic scenes.

The main contributions of this work are as follows.

- This paper proposes a new multi-modal traffic scene reconstruction framework, which also utilizes the 3D model of the image sequence and road to construct the traffic scene by two registration steps. Different from

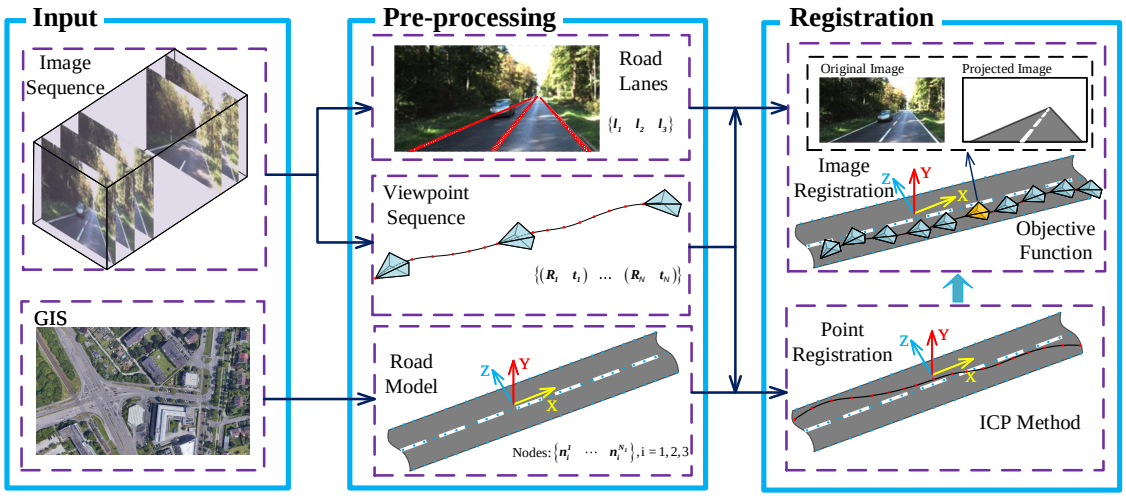


Fig. 1: Overview of the proposed pipeline

existing 3D reconstruction methods, the registration is implemented through estimating the pose of roadways during a coarse-to-fine process.

- In the registration stage, a new objective function is proposed to guarantee the consistency between the reconstructed 3D model and the captured road image sequence.

II. RELATED WORK

In recent years, there are many image-based model methods for traffic scene reconstruction, thanks to the rapid advancements in computer vision techniques [10] and remote sensing research [11]–[13]. The most popular methods include structure from motion (SfM), multi-view stereo (MVS) [14], Simultaneous localization and mapping (SLAM), etc. SfM methods can construct a large scene from the images in network. In general, the bundle adjustment is employed to optimize the recovered 3D points in the post-processing stage. The successful examples are that Agarwal [15] reconstructed the Rome city from the community photo collections(Flickr); Heinly [16] utilized 100 million image from YAHOO to model the world. Thus, the advantage of these methods is flexible on input images, while the limitation is that the reconstructed 3D points are so sparse to contain any solid geometry. In order to overcome the sparsity problem. Many researchers proposed the MVS methods [17], which aim to find best match for all pixels. Thus, in generally, the feature correspondences of any pixel can be obtained by searching in the reference image. In addition, some researchers study the vision-based SLAM algorithms to acquire the accurate visual odometry and traffic scene model simultaneously. For example, Andreas Geiger [18] utilize a circle of four images, i.e.left and right images of two consecutive frames, to obtain robust feature correspondences for visual odometry and scene reconstruction. The main advantage of SLAM method is real time.

However, all mentioned above can be summarized as the bottom-to-up methods and the reconstructed scene lacks the semantic information. Namely, it is hard to distinguish the road, building and vehicles etc. from the scene. In order to make semantic scene, Cornelis [7] added the detection algorithm to the SfM framework and the reconstructed urban city includes roads, building and movable objects. Furthermore, some researchers aim to model the traffic scene for vehicle simulation test. In [9], the 'floor-wall' model, which divided a traffic scene into road plane, right wall, left wall and back wall, was employed to construct scenes for unmanned vehicle test.

The reminder of this paper is organized as follows: Section III provides an overviews of the proposed method, including the problem definition and the proposed framework. Section IV describes preprocessing details about the multi-model data (GIS, image sequence). Multi-model data inputs are combined and registered in traffic scenes in two stages, as described in section V. Section VI presents the experimental results and corresponding analyses. Finally the conclusions and potential future work are provided in section VII.

III. OVERVIEWS

A. Problem Formulation

As is shown in Fig 1, Given the image sequence, camera intrinsic matrix $\mathbf{K}$ and the GIS information of the corresponding road, the traffic scene containing images sequence and road can be constructed. In this problem, the road model, the location, orientation of the road and each image are unknown. Thus, in order to construct the traffic scene from image sequence and GIS information, the road model must be generated. In addition, location (i.e. translation $\mathbf{t}_m$) and orientation (i.e. rotation $\mathbf{R}_m$) of the road model, the location (i.e. translation $\mathbf{t}_i$) and pose (i.e. rotation $\mathbf{R}_i$) of the ith image must be accurately estimated.

B. the Pipeline of the Proposed Method

To solve the problem mentioned in Sec.III.(A), our method is proposed and shown in the Fig 1. In the pre-processing stage, our method first makes the road model from the GIS information(the details in Sec.IV.(A)), In parallel, combination of the image sequence and intrinsic matrix $\mathbf{K}$ can

recover the rotation $\mathbf{R}_i$ and translation $\mathbf{t}_i$ (in Sec.IV.(B)). To estimate the location $\mathbf{t}_m$ and orientation $\mathbf{R}_m$ of the road model in the registration stage, the solution includes two steps: point registration and image registration. Point registration is to utilize the iterative closest point (ICP) method [19] [20] to estimate the $\mathbf{R}_m$ and $\mathbf{t}_m$ coarsely (in Sec.V.(A)). Finally, in the image registration, the $\mathbf{R}_m$ and $\mathbf{t}_m$ (i.e. 6DoF) of the road model is refined by optimizing the objective function, which reduces the pixel errors between the image and reprojected region of the road model (in Sec.V.(C)).

IV. MULTI-MODEL DATA PREPROCESSING

In this paper, multiple modalities of sensors are exploited to capture the road and traffic information. Due to the limitation of such sensors [21], [22], it is challenging to match and perfectly calibrate such multi-model data directly. Therefore, it is necessary to pre-process the multi-modal data before combining them and registering them in reconstructed traffic scenes. Specifically, two data modalities (i.e., GIS and image sequence) are employed in this paper. This preprocessing includes three stages: 3D road model generation from GIS information, poses of viewpoints (odometry) estimation from image sequences and lanes estimation from images.

A. Road Model

The 3D road model utilized in this paper is inspired by the corridor model [8] except for the right and left walls. Especially, we only have the road surface. To acquire the road model from the Google Earth, we first localize the road by start and end landmarks (the yellow landmarks in Fig. 2), Then the lanes of the road are labeled by nodes in the Google Earth (the white lines in Fig. 2). After exporting the labeled files, the nodes of lanes are converted from the WGS84 [23] to ENU coordinate. Concretely, Given the longitude λ, latitude ϕ and height h, the point can be converted into ECEF coordinate by Eq. (1). Without loss of generality, the point $p_0(\lambda_0, \phi_0, h_0)$ is regarded as the origin of ENU coordinate and any point relative to the ENU coordinate can be calculated by Eq. (2).

$$\begin{aligned} x &= (a/\chi + h)cos(\phi)cos(\lambda) \\ y &= (a/\chi + h)cos(\phi)sin(\lambda) \\ z &= (a(1-e^2)/\chi + h)cos(\phi)sin(\lambda) \end{aligned} \tag{1}$$

where $\chi = \sqrt{1 - e^2 sin^2(\phi)}$, $a = 6378137$, $e^2 \approx 6.69 \times 10^{-3}$

$$\begin{aligned} p^i_{ENU} &= \mathbf{R}^{ENU}_{ECEF} \times (p^i_{ECEF} - p^0_{ECEF}) \\ \mathbf{R}^{ENU}_{ECEF} &= \begin{bmatrix} -sin(\lambda_0) & cos(\lambda_0) & 0 \\ -sin(\phi_0)cos(\lambda_0) & -sin(\phi_0)sin(\lambda_0) & cos(\phi_0) \\ cos(\phi_0)cos(\lambda_0) & cos(\phi_0)sin(\lambda_0) & sin(\phi_0) \end{bmatrix} \end{aligned} \tag{2}$$

B. Odometry

In this part, given the image sequence and the intrinsic matrix $\mathbf{K}$, the location $\mathbf{R}_i$ and orientation $\mathbf{t}_i$ of the ith image are estimated. This paper utilizes the ORB feature [24] based SLAM method (i.e.ORB-SLAM2 [25]) to estimate the viewpoints of image sequences. To extract the ORB feature, we set six-level pyramid and extract one thousand amount of features [24] for an image.

C. Lane

In Fig.1, the lanes are extracted from images. In order to obtain accurate lanes position, we label the straight lanes manually in some frames. Finally, the line is linearly fitted by the labeled points.

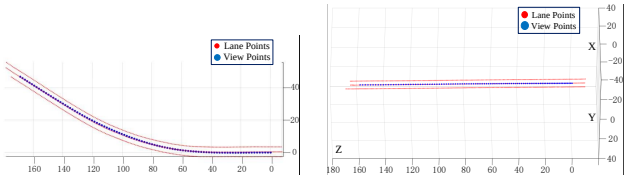
V. REGISTRATION

Since the viewpoints and the road model is measured in their own 3D coordinate, we want to register both the viewpoints and road model in a coordinate. As is mentioned in Sec. III(B), the process of registration includes two stages: point registration and image sequence registration.

A. Point Registration

In this stage, we utilize the nodes of a road lane in the model to represent the road. The viewpoints of image sequences can be also regarded as the point cloud. Thus, the iterative closest point (ICP) algorithm [19], [20] can be employed to match the road and viewpoints. Then the viewpoints can be registered in the road model. Fig. 3 shows a point registration example by 'Point-to-Point' ICP algorithm [19], [20]. It is obvious that the viewpoints can match the one of lanes in both curve and straight roads. However, the viewpoints do not register in the accurate position but in the range of the road surface.

Fig. 2: the road model from the Google Earth



(a) curve road (b) straight road

Fig. 3: Illustration of point registration

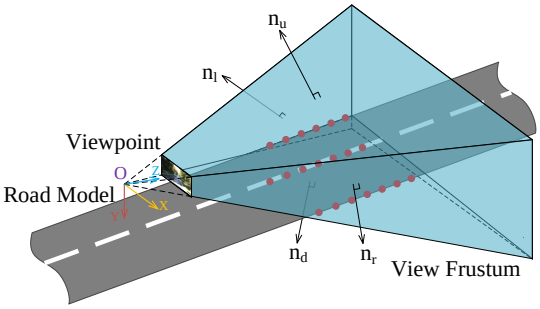


Fig. 4: a figure of selecting projected points in the road

B. Projected Point

Before image sequence registration, the suitable nodes of the road model need to be selected for their corresponding viewpoints. The principle to select nodes of road model for each viewpoint is that the road nodes can be seen in the viewpoint. In other words, the points must locate in the region of a view frustum. As is shown in the blue region of Fig. 4, a view frustum can be determined by five planes (up plane, down plane, left plane, right plane, image plane). Given the intrinsic parameters $\mathbf{K}$ and image size, the planes can be expressed in Eq. (3), where $\boldsymbol{p}_{i,i\in\{ul,dl,dr,ur\}}$ are the pixel coordinate of up-left, down-left, down-right, up-right point in the image and $\boldsymbol{n}_{i,i\in\{d,l,r,u\}}$ are the normal vectors of four planes (shown in Fig. 4). Thus, the region of view frustum $D(f)$ can be expressed in Eq. (4).

$$\begin{aligned} & z-f=0; \mathbf{n}_i^T \boldsymbol{p}=0; i \in\{u, d, l, r\} \\ & where,\ \boldsymbol{n}_l=(\mathbf{K}^{-1}\boldsymbol{p}_{ul}) \times (\mathbf{K}^{-1}\boldsymbol{p}_{dl}); \\ & \boldsymbol{n}_u=(\mathbf{K}^{-1}\boldsymbol{p}_{ul}) \times ((\mathbf{K}^{-1}\boldsymbol{p}_{ur})) \\ & \boldsymbol{n}_r=(\mathbf{K}^{-1}\boldsymbol{p}_{dr}) \times ((\mathbf{K}^{-1}\boldsymbol{p}_{ur})); \\ & \boldsymbol{n}_d=(\mathbf{K}^{-1}\boldsymbol{p}_{dr}) \times ((\mathbf{K}^{-1}\boldsymbol{p}_{dl}) \end{aligned} \tag{3}$$

$$D(f)=\{\boldsymbol{p} | \boldsymbol{n}_i^T \boldsymbol{p}>0; z-f>0; i \in\{l, r, u, d\}\} \tag{4}$$

In terms of selecting the suitable points for ith viewpoint. all the nodes of road model first are converted into the ith camera coordinate by the rotation matrix $\mathbf{R}_i$ and transform vector $\mathbf{t}_i$. Then the nodes satisfying the Eq. (4) are in the view frustum $D(f)$ (red points in Fig. 4). Finally, we select top ten nodes near to ith viewpoint for each lane.

C. Image Sequence Registration

In this stage, the objective function is designed to achieve the image sequence registration. After point registration, the road model still does not consist with the image sequence strictly. Thus, we regard the road model as the rigid object, and the translation can make the road model complies the image sequence.

We have ith rotation matrix, translation vector $\mathbf{R}_i, \mathbf{t}_i$ and intrinsic matrix $\mathbf{K}$, an amount of N lane lines $\{l_i^1, l_i^2, ..., l_i^N\}$ in ith image (the number of lines depends on the image) and the number of Q nodes $\{P_i^{j_1}, P_i^{j_2}, ...P_i^{j_k}, ..., P_i^{j_Q}\}$ of the road model for the jth lane of the ith image in the model coordinate. Assume that the road model can consist with the image sequence after rotation matrix $\mathbf{R}_m$ and translation vector $\boldsymbol{t}_m$

In terms of the ith viewpoint, the nodes of the jth lane can be projected into the image according to the following equation.

$$\boldsymbol{p}_i^{jk}=\mathbf{K}(\mathbf{R}_i(\mathbf{R}_m \boldsymbol{P}_i^{jk}+\mathbf{t}_r)+\mathbf{t}_i) \tag{5}$$

However, the $\boldsymbol{p}_i^{jk}$ is in homogeneous format. Thus, the vector $\mathbf{m}=\begin{bmatrix}0 & 0 & 1\end{bmatrix}$ is employed to extract the z value of the projected point p_i^{jk}, and the true pixel position is $\frac{\boldsymbol{p}_i^{jk}}{(\mathbf{m}\boldsymbol{p}_i^{jk})}$. Then, the Euclidean distance between the projected point and lane is calculated by the following equation.

$$D_i^{jk}=\|(\boldsymbol{l}_i^j)^T \boldsymbol{p}_i^{jk}\|_2 \tag{6}$$

In terms of all viewpoints, summation of all the Euclidean distance is the objective function.

$$E(\mathbf{R}_m, \boldsymbol{t}_m)=\sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{Q} D_i^{jk} \tag{7}$$

As is known, the matrix $\mathbf{R}_m$ has only three degree of freedom. Thus, the matrix $\mathbf{R}_m$ are replaced by the rotation axis $\boldsymbol{v}$ and angle θ, namely Rodrigues [26] equation $\mathbf{R}_m = cos(\theta)\mathbf{I} + sin(\theta)[\boldsymbol{v}]_\times + (1-cos(\theta))\boldsymbol{v}\boldsymbol{v}^T$. Thus, the final format of the objective function is shown in Eq. (8), where the rotation axis $\boldsymbol{v}$ is the unit vector; the axis degree θ is in range of $-\pi/3$ and $\pi/3$ radian; and the components of the translation vector $\boldsymbol{t}_m$, i.e. $t_{m,x}$, $t_{m,y}$, $t_{m,z}$, are in range of -5 and 5 meter.

$$\begin{aligned} E(\mathbf{v}, \theta, \boldsymbol{t}_m)=\sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{Q} D_i^{jk}(\mathbf{v}, \theta, \boldsymbol{t}_m) \\ s.t.\ \|\boldsymbol{v}\|_2=1 \\ -\pi/3<\theta<\pi/3 \\ -5<\mathbf{t}_{m,i}<5,\ i \in x, y, z \end{aligned} \tag{8}$$

For optimization, it is hard to get close-form solution for the objective function (Eq. (8)). Thus, the gradient descent method based iterative scheme is employed for optimization. Meanwhile, a good initial value and suitable step size is must. In actual, the initial value of the rotation axis $\boldsymbol{v}$, rotation angle θ and translation vector $\boldsymbol{t}_m$ are the normal vector of the road, 0 degree and zero vector respectively. The interior algorithm [27] are exploited for iterative updating. To avoid obtaining the local minimum by single initial value, we exploit the scatter-search mechanism [28] to acquire multiple initial values in the scope of constraints. Through comparing all the local solutions, we can find the global minimum.

VI. EXPERIMENT

A. Experiment Preparation

1) Dataset: The raw KITTI dataset [6] are exploited and processed for evaluating our method. In order to test our method in all aspects, we select four groups of data, i.e. city_09_26_14, city_09_26_96, road_09_26_15, road_09_26_28, and each of the image sequences has 227, 240, 240, 170

frames respectively. As is shown in Fig. 5, the four group image sequence contains both curved and straight roads. For each group, the image sequence, and intrinsic matrix of camera has been extracted.

2) Metrics: In this paper, we want to evaluate our method in the 2D space. Concretely, If the 3D road model is located in accurate position, the reprojected region of 3D road model is strictly the same with that of the image, namely the reprojection errors between road model and image are as small as possible. Furthermore, the reprojection errors of all frames of the image sequence is small, when the 3D road registers accurately. Thus, Similar to [29], we utilize the intersection over union (shown in Eq. (9)) between reprojected road model and the road region of each image.

$$IoU = \frac{S(reprojection) \bigcap S(Groundtruth)}{S(reprojection) \bigcup S(Groundtruth)} \tag{9}$$

B. Results and Analyses

After the image sequences and road model registration, the road and viewpoints of sequences are integrated into a whole scene. Fig. 5 shows point structure of the traffic scenes in the vertical view. In Fig. 5, the red, blue and black points indicate the road boundaries, lane and viewpoints respectively. Obviously, the viewpoints are in the scope between right road boundary and central lane, which indicates that the vehicle keeps right for driving and accords with the reality.

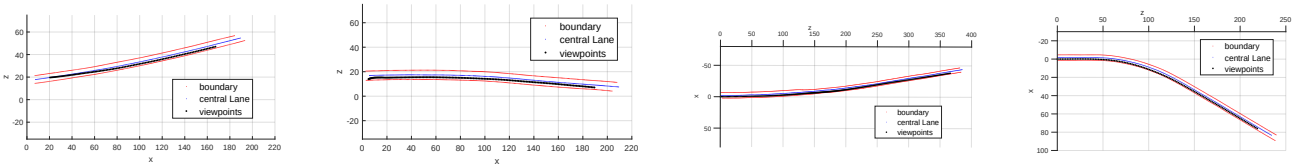
The IoU scores of image sequences are shown in the Fig. 6, where the red line indicates that the IoU score varies with the frame, and the blue line represent average IoU of all frames. The IoU scores in Fig. 6 (a),(b),(c) and (d) approximately range from 0.88 to 0.98, from 0.89 to 0.99, from 0.88 to 0.99, and from 0.905 to 0.991 respectively. The average IoU scores of each sequence are 0.938, 0.951, 0.949, 0.958 respectively. As a result, the reprojected region of the 3D road model basically accords with the road regions of the image sequences after registration. In addition, the proposed method has good performance in both straight and curve roads.

However, as is shown in Fig. 6, the IoU scores of some successive frames are much lower than the average score. In general, it appears that IoU score gradually descends and then ascends. The typical examples are shown in Fig. 6(b) (from the 70th to120 frames), in Fig. 6(c) (from the 70th to the 100th frames) and in Fig. 6(d) (from the 40th to 60th frames). What kind of factor influences the performance of the proposed method. In terms of 'city09_26_0096' sequence, the slant angle of the road changes from 70th to 120th frames. For 'road09_26_0015' sequence, the left boundary of road suddenly varies in the Fig. 5(c), which brings about the width of the road narrow. Meanwhile, the change of the width of the road results in the low IoU scores from 40th to 60th frames in Fig. 6(c). As to 'road 09_26_0028' sequence, the viewpoints from 40th to 60th frames are located in the turning of the road, which results in the low IoU scores from 40th to 60th frames in Fig. 6(d). Thus, it is obvious that the IoU scores descend gradually when the road model does not consist with the real road strictly, especially in some suddenly changing sections.

In addition, we observe that the IoU scores vary dramatically between adjacent frames, which leads to the serrated curves in Fig. 6. We speculate that this is caused by the noisy estimations of poses of viewpoints by the SLAM [25] method. The statistics of all four image sequences are summarized in Table I, with percentage of frames achieving predefined IoU scores. From Fig. 5(d) and last row of Table I, it is evident that the proposed method is robust to both straight and curvy roadways.

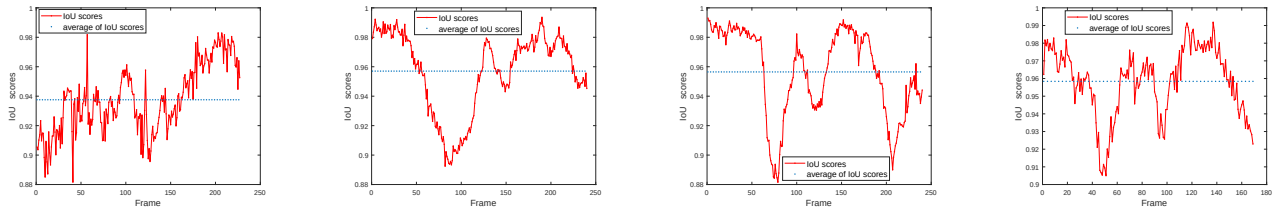
VII. CONCLUSION AND FUTURE WORK

In this paper, we propose a new two-stage traffic scene reconstruction framework that exploits multi-model information (image sequence, GIS). In the first stage, a 3D road



(a) city 09_26_0014 (b) city 09_26_0096 (c) road 09_26_0015 (d) road 09_26_0028

Fig. 5: the point structure of the traffic scene in the vertical view



(a) city 09_26_0014 (b) city 09_26_0096 (c) road 09_26_0015 (d) road 09_26_0028

Fig. 6: the IoU scores of image sequences

TABLE I: the statistical results of IoU scores

Image Sequence	the number of the frames	IoU scores		
		>90%	>94%	>98%
City09_26_0014	227	96.04%	43.61%	3.08%
City09_26_0096	240	96.67%	77.50%	19.58%
Road09_26_0015	240	94.58%	72.58%	34.17%
Road09_26_0028	170	100.00%	82.35%	11.18%

model and viewpoints are obtained from the GIS information and image sequences, respectively. In the second stage, the iterative closest point (ICP) algorithm is utilized for estimation of the pose of roads. Subsequently, the pose is refined by the proposed new objective function and the traffic scene reconstruction is achieved. Quantitative experiments demonstrate that the IoU scores are almost consistently higher than 0.9 and the reconstructed scenes correlate highly with the image sequences.

For potential future work, we are planning to investigate in three aspects, including a more robust automatic lane detection algorithm, the few failure cases of the proposed framework which typically involves abrupt turns in roads and even more complex traffic scene such as urban road intersections.

ACKNOWLEDGMENT

This research has been supported by National Natural Science Foundation of China (NSFC) under Grant No.91520301.

REFERENCES

[1] L. Ran, Y. Zhang, W. Wei, and Q. Zhang, "A hyperspectral image classification framework with spatial pixel pair features," *Sensors*, vol. 17, no. 10, p. 2421, 2017.

[2] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA*, 2-7 February, 2018.

[3] L. Ran, Y. Zhang, Q. Zhang, and T. Yang, "Convolutional neural network-based robot navigation using uncalibrated spherical images," *Sensors*, vol. 17, no. 6, p. 1341, 2017.

[4] J. Zang, L. Wang, Z. Liu, Q. Zhang, Z. Niu, G. Hua, and N. Zheng, "Attention-based temporal weighted convolutional neural network for action recognition," in *Proceedings of the 14th IFIP WG 12.5 International Conference on Artificial Intelligence Applications and Innovations, AIAI 2018, Rhodes,Greece*, 2018.

[5] C. Zhang, Y. Liu, D. Zhao, and Y. Su, "Roadview: A traffic scene simulator for autonomous vehicle simulation testing," in *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*. IEEE, 2014, pp. 1160–1165.

[6] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[7] N. Cornelis, B. Leibe, K. Cornelis, and L. Van Gool, "3d urban scene modeling integrating recognition and reconstruction," *International Journal of Computer Vision*, vol. 78, no. 2-3, pp. 121–141, 2008.

[8] Y. Li, Y. Liu, C. Zhang, D. Zhao, and N. Zheng, "The floor-wall traffic scenes construction for unmanned vehicle simulation evaluation," in *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*. IEEE, 2014, pp. 1726–1731.

[9] Y. Li, Y. Liu, Y. Su, G. Hua, and N. Zheng, "Three-dimensional traffic scenes simulation from road image sequences," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 1121–1134, 2016.

[10] Q. Zhang, G. Hua, W. Liu, Z. Liu, and Z. Zhang, "Auxiliary training information assisted visual recognition," *IPSJ Trans. Comput. Vis. and Appl.*, vol. 7, pp. 138–150, 2015.

[11] Q. Zhang, H. Abeida, M. Xue, W. Rowe, and J. Li, "Fast implementation of sparse iterative covariance-based estimation for array processing," in *Proceedings of the Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA*, 6-9 November, 2011, pp. 2031–2035.

[12] H. Abeida, Q. Zhang, J. Li, and N. Merabtine, "Iterative sparse asymptotic minimum variance based approaches for array processing," *IEEE Trans. Signal Process.*, vol. 61, no. 4, pp. 933–944, 2013.

[13] Q. Zhang, H. Abeida, M. Xue, W. Rowe, and J. Li, "Fast implementation of sparse iterative covariance-based estimation for source localization," *J. Acoust. Soc*, vol. 131, no. 2, pp. 1249–1259, 2012.

[14] P. Musialski, P. Wonka, D. G. Aliaga, M. Wimmer, L. v. Gool, and W. Purgathofer, "A survey of urban reconstruction," in *Computer graphics forum*, vol. 32, no. 6. Wiley Online Library, 2013, pp. 146–177.

[15] S. Agarwal, Y. Furukawa, N. Snavely, B. Curless, S. M. Seitz, and R. Szeliski, "Reconstructing rome," *Computer*, vol. 43, no. 6, pp. 40–47, 2010.

[16] J. Heinly, J. L. Schonberger, E. Dunn, and J.-M. Frahm, "Reconstructing the world* in six days*(as captured by the yahoo 100 million image dataset)," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3287–3295.

[17] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Towards internet-scale multi-view stereo," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1434–1441.

[18] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *Intelligent Vehicles Symposium (IV), 2011 IEEE*. Ieee, 2011, pp. 963–968.

[19] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," *Image and vision computing*, vol. 10, no. 3, pp. 145–155, 1992.

[20] P. J. Besl, N. D. McKay *et al.*, "A method for registration of 3-d shapes," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[21] Q. Zhang and G. Hua, "Multi-view visual recognition of imperfect testing data," in *Proceedings of the ACM International Conference on Multimedia, Brisbane, Australia*, 26-30 October, 2015, pp. 561–570.

[22] Q. Zhang, G. Hua, W. Liu, Z. Liu, and Z. Zhang, "Can visual recognition benefit from auxiliary information in training?" in *Proceedings of the Asian Conference on Computer Vision, Singapore*, 1-5 November, 2014, pp. 65–80.

[23] S. Drake, "Converting gps coordinates to navigation coordinates (enu)," 2002.

[24] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *Computer Vision (ICCV), 2011 IEEE international conference on*. IEEE, 2011, pp. 2564–2571.

[25] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[26] W. contributors, "Rotation matrix — wikipedia, the free encyclopedia," 2017, [Online; accessed 17-January-2018]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Rotation_matrix&oldid=816278453

[27] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, "An interior algorithm for nonlinear optimization that combines line search and trust region steps," *Mathematical programming*, vol. 107, no. 3, pp. 391–408, 2006.

[28] Z. Ugray, L. Lasdon, J. Plummer, F. Glover, J. Kelly, and R. Martí, "Scatter search and local nlp solvers: A multistart framework for global optimization," *INFORMS Journal on Computing*, vol. 19, no. 3, pp. 328–340, 2007.

[29] H. Tjaden, U. Schwanecke, and E. Schömer, "Real-time monocular pose estimation of 3d objects using temporally consistent local color histograms," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 124–132.