2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Supervised Articulation Angles Estimation for Multi-Articulated Vehicles Based on Panoramic Camera System

Weimin Liu[1], Wenjun Wang[2*], Zhaocong Sun[3]

***Abstract*— Articulation angle plays a significant role in determining the motion of a complex dynamic system such as a multi-articulated vehicle. By engineering practice, articulation angles are measured using mechanical angle sensors that are delicate to physical damage. To overcome this problem, this study proposed a supervised articulation angle estimation method based on the panoramic camera system of multi-articulated vehicles. By constructing neural network that takes images of surrounding environment captured by spatially adjacent cameras as input, and takes temporal dependency as well as data imbalanced distribution into consideration, we show that the proposed vision-only method could make accurate estimations either on collected dataset or field experiment. Results of our experiments verified the validity and feasibility of the proposed method in playing as an alternative to mechanical angle sensors without bringing additional hardware setting expenses.**

I. Introduction

Multi-articulated vehicle, such as autonomous rail rapid transit (ART), representing a brand new pattern of transporation means, has now become a significant and popular transportation means in many metropolitans of China such as Shanghai and Nanjing. In comparison with traditional passenger vehicles, multi-articulated vehicles are composed of multiple carbodies articulated by hinge joints and therefore have the advantage of high transport efficiency. However, as a significantly more complex dynamic system, multi-articulated vehicles are confronted with more challenges such as being more susceptible to unstable motions [1]. In the last decade, we've seen researchers sparing their efforts on addressing these potential instabilities, where articulation angle has always been a necessary and inevitable variable in developing automated driving systems [2] [3], as it composes the most distinguishable and decisive characteristic of the dynamics of multi-articulated vehicles in against to that of passenger vehicles. Yet, the measurement or the estimation of the articulation angle could be challenging.

By engineering practice, angle sensors are commonly installed in the vicinity of hinge points to realize real-time measurements. However, one the of most common problems is that angle sensors are prone and delicate to physical damage, which results in inconvenience in replacing them periodically, as coupling and decoupling of articulated vehicles could be strenuous. Therefore, we seek to utilize a vision-based method as a substitute for mechanical and physical measurements. Inspired by works of self-supervised monocular depth estimation and camera ego-motion estimation [4] [5] [6], where PoseNet takes two temporally adjacent images taken by the same camera to estimate temporal pose transformation, we extended this paradigm to spatial PoseNet which takes spatially adjacent images captured by two cameras at the same time to estimate spatial pose transformation and further recover articulation angles. In addition, as multi-articulated vehicle consists of multiple carbodies, we aim to realize full articulation estimations based on surround panoramic cameras mounted.

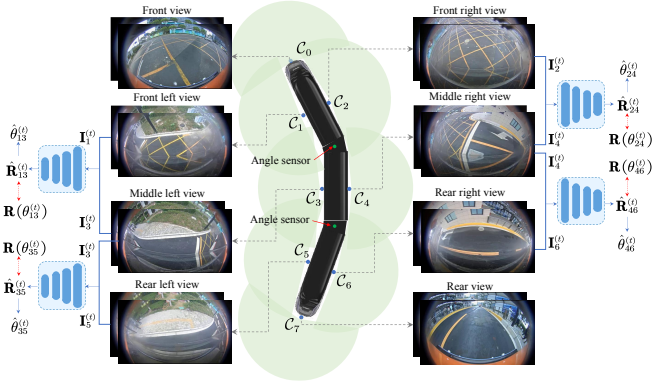

Fig. 1. Overview of supervised articulation angle estimation for a three-carbody-articulated vehicle based on panoramic camera system: Network takes two images captured by spatially adjacent cameras and outputs their pose transformation, from where estimated articulation angles are recovered.

Panoramic camera system comprises multiple cameras mounted on each side of the vehicle, creating an omni-directional, multi-perspective camera array [7]. Mature usage of panoramic camera systems can be found in the applications of panoramic view on passenger vehicles to assist drivers with automatic parking or near-field blind spot sensing. By academic research, panoramic camera system has been studied in depth to realize for example bird's eye view (BEV) perception [8] [9], where overlapping between cameras provides spatial transformation information. Our work aims to leverage this and extend it to articulated rigid bodies to realize the estimation of articulation angles.

Fig. 1 shows an overview of our proposed supervised articulation angle estimation algorithm on a three-carbody-articulated vehicle. Corresponding panoramic camera system consists of eight fisheye cameras (denoted as $\mathcal{C}_0 \sim \mathcal{C}_7$) in field-of-view (FoV) of 190° to realize surround environment perception, where cameras mounted on the side of the

[1]Weimin Liu is with School of vehicle and Mobility, Tsinghua University, Beijing, China lwm23@mails.tsinghua.edu.cn
[2*]Wenjun Wang (corresponding author) is with School of vehicle and Mobility, Tsinghua University, Beijing, China wangxiaowenjun@tsinghua.edu.cn
[3]Zhaocong Sun is with School of vehicle and Mobility, Tsinghua University, Beijing, China szc19@mails.tsinghua.edu.cn

vehicle were utilized to capture the overlapping relationship between adjacent carbody and further estimate articulation angles. Dataset was collected including sequential images captured by side cameras and temporally aligned groundtruth articulation angles recorded by mechanical angle sensors. We trained and tested our network on the datasets, where the results verified the validity of our proposed method.

The contributions of our work are as follows:

- We proposed a network ArticuNet to estimate the spatial pose transformation with images of surrounding environment taken by spatially adjacent panoramic cameras, where articulation angles can be subsequently decomposed from the estimated spatial poses.
- Dataset was collected based on the panoramic camera system settings and used for network training. By implementing experiments, we show estimation errors could be reduced by leveraging redundant camera pairs.
- To deal with imbalanced distribution of articulation angle observations, Gaussian-fitted probability density function (PDF) was adopted as weighting coefficient of supervision losses to penalize sparsely distributed cases.
- The proposed method could be regarded as an alternative to mechanical angle sensor in future development of vehicle dynamic control algorithms.

II. Related Works

Non-learning-based methods. Scholars have attempted to leverage non-mechanical contact sensors for the estimation of articulation angles. Olutomilayo *et al.* [10] [1] uses 2D pointclouds collected from radars installed in the taillight fixtures of a trailer-coupled truck to estimate the angle between tractor and trailer. Saxe *et al.* [11] uses a rear trailer-facing camera and the parallel tracking and mapping (PTAM) image processing algorithm to realize articulation angle estimation. Though showing valid estimations, the vision-based algorithm is sensitive to the number and distribution of features observed in the scene. Peng *et al.* [7] developed an articulated multi-perspective camera (AMPC) system for the estimation of articulation angle through temporally aligning keypoints captured by two non-overlapping monoculars individually mounted at the very opposite end of each carbody. The results of their study show accurate motion estimations by additionally introducing Ackermann motion constraints. Apart from using natural image features for estimation, Fuchs *et al.* [12] [13] used artificial markers to estimate articulation angle based on geometric pose estimation. The proposed method can also be applied to estimate pitch and roll angles. However, the dependency of artificial markers limits the algorithm from real practice and application.

Learning-based methods. As computer vision constitutes a crucial part of deep-learning studies, we've also seen research combining these two in solving articulation angle estimations. Atoum *et al.* [14] proposed an automated vision-based deep-learning system for autonomous vehicle self-backing towards a trailer. Kaci *et al.* [15] proposed a convolution neueral network (CNN) architecture to process the image containing the hinge component to estimate articulation angle. However, this method can not be applied to multi-articulated vehicles whose gangway makes the hinge component not directly visible by optical sensors.

In summary, empirical studies could realize articulation angle or hinge angle estimations on some extent, yet algorithms in most of these studies cannot be directly applied to multi-articulated vehicles due to reasons such as high dependency on visible articulation components or artificial markers. Therefore, in this study, we proposed a supervised deep-learning approach based on multi-perspective cameras of the panoramic camera system mounted on the side of carbodies to estimate articulation angles by learning overlapping between spatially adjacent cameras without being highly conditioned on ideal prior settings.

III. Methodology

A. Spatial Pose Transformation

Let $\mathbf{T}_{ji} = [\mathbf{R}_{ji}|\mathbf{t}_{ji}]$ be the spatial pose transformation variables from $\mathcal{C}_i$ to $\mathcal{C}_j$ such that a 3D point $\mathbf{p}_i$ in $\mathcal{C}_i$ can be transformed to $\mathcal{C}_j$ following $\mathbf{p}_j = \mathbf{R}_{ji}\mathbf{p}_i + \mathbf{t}_{ji}$, where $\mathbf{R}_{ji} \in \mathrm{SO}(3)$ is rotation and $\mathbf{t}_{ji} \in \mathbb{R}^3$ indicates translation. The calculation of rotation matrix is given as follows,

$$\mathbf{R}_{ji} = {}^{x}\mathbf{R}_{j}^{T}(\gamma) \cdot {}^{y}\mathbf{R}_{ji}(\theta) \cdot {}^{x}\mathbf{R}_{i}(\gamma), \tag{1}$$

where ${}^{x}\mathbf{R}_{i}(\gamma)$ indicates the rotation concerning installation angle γ with pure rotation of x-axis of $\mathcal{C}_i$ (see Fig. 3), and ${}^{y}\mathbf{R}_{ji}(\theta)$ represents the rotation from $\mathcal{C}_i$ to $\mathcal{C}_j$ in angle of θ over vehicle planar motion surface (around y-axis of camera). ${}^{x}\mathbf{R}_{i}(\gamma)$ and ${}^{y}\mathbf{R}_{ji}(\theta)$ can be mathematically expressed as follows in (2) and (3), respectively,

$${}^{x}\mathbf{R}_{i}(\gamma) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\pi/2-\gamma) & \sin(\pi/2-\gamma) \\ 0 & -\sin(\pi/2-\gamma) & \cos(\pi/2-\gamma) \end{bmatrix}, \tag{2}$$

$${}^{y}\mathbf{R}_{ji}(\theta) = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix}. \tag{3}$$

Translation deduction needs up-to-scale measurements of the physical vehicle, as it represents the spatial displacements from one camera to the other. Let $\mathbf{t}_w = [0, 0, s_w]^T$ denote the lateral displacement from one to the other side of the carbody, and $\mathbf{t}_l = [s_l, 0, 0]^T$ denote the longitudinal displacement from camera to hinge point. s_w is equal to the width of carbody and s_l implies the longitudinal side distance from camera installation to hinge point. By simple pose transformation and mathematical deduction, the translation from $\mathcal{C}_i$ to $\mathcal{C}_j$ can be written as follows,

$$\mathbf{t}_{ji} = {}^{x}\mathbf{R}_{j}^{T}(\gamma) \cdot \left[{}^{y}\mathbf{R}_{ji}(\theta) \cdot \left(\frac{\mathbf{t}_w}{2} - \mathbf{t}_l \right) - \left(\frac{\mathbf{t}_w}{2} + \mathbf{t}_l \right) \right]. \tag{4}$$

As calculation of translation $\mathbf{t}_{ji}$ given in (4) also includes component of ${}^{y}\mathbf{R}_{ji}(\theta)$, we thus only use $\mathbf{R}_{ji}$ as estimation objective to avoid over-parametering.

The overall spatial pose transformations of panoramic camera system on a three-carbody-articulated vehicle are

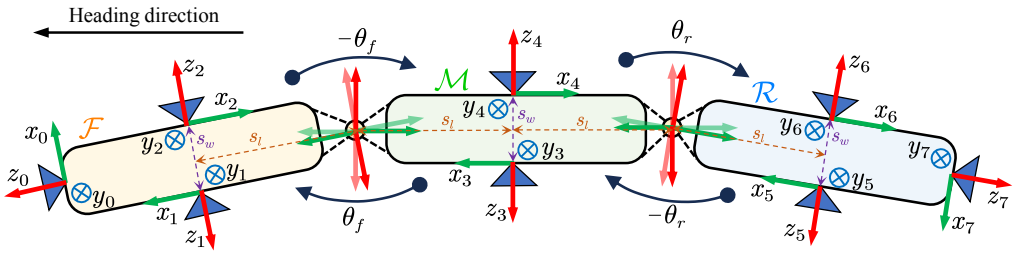

Fig. 2. Spatial pose transformation between adjacent fisheye cameras and camera locations on a three-carbody-articulated vehicle (xoz plain of camera have been rotated to vehicle motion plain with compensation of camera installation angle γ; articulation angle θ to be positive (negative) when front or rear carbody rotates towards left (right) side of middle carbody: for instance, in above figure, $\theta_f > 0$ and $\theta_r < 0$.)

Fig. 3. Installation angle γ (red arrow indicates z-axis of camera).

shown in Fig. 2, where camera pairs $\mathcal{C}_4\&\mathcal{C}_2$ and $\mathcal{C}_1\&\mathcal{C}_3$ are responsible for the estimation of front articulation angle θ_f, and $\mathcal{C}_6\&\mathcal{C}_4$ and $\mathcal{C}_3\&\mathcal{C}_5$ are responsible for the estimation of rear articulation angle θ_r.

As one articulation angle could be estimated by two camera pairs that have different ranges of overlapping areas, to prevent two camera pairs from being labeled with the same spatial rotation transformation as groundtruth, we manually set the groundtruth articulation angle of one side of two articulated carbodies to maintain the original value for spatial rotation calculation, yet the other side with the inverse signed one, which concretely, substituting θ with $-\theta$ (see Fig. 2).

B. Network Architecture

We propose network **ArticuNet** for supervised articulation angle estimation. The network consists of three components, including pose encoder, single-layer ConvLSTM [16] and pose decoder. Overall network architecture is shown in Fig. 4. Inspired by self-supervised simultaneous monocular depth estimation and camera pose estimation, we leveraged a similar pattern of pose encoder for images captured by spatially adjacent cameras instead of two images captured by the same camera yet temporally adjacent. Single-layer ConvLSTM is used to establish and compensate temporal dependency considering articulation angles (or in other words, spatial pose) temporally adjacent would not change abruptly as being constrained by vehicle dynamics and physical motions. The output of each ConvLSTM cell is fed to pose decoder for the estimation of spatial poses. The overall process can be written as follows in (5)∼(7),

$$\mathbf{x}^{(t)} = \texttt{PEnc}\left(\left[\mathbf{I}_j^{(t)}, \mathbf{I}_i^{(t)}\right]\right), \tag{5}$$

$$\mathbf{h}^{(t)} = \texttt{ConvLSTM}\left(\mathbf{h}^{(t-1)}, \mathbf{c}^{(t-1)}, \mathbf{x}^{(t)}\right), \tag{6}$$

$$\hat{\mathbf{R}}_{ji}^{(t)} = \texttt{PDec}\left(\mathbf{h}^{(t)}\right), \tag{7}$$

where PEnc indicates pose encoder, $\mathbf{I}_j^{(t)} \in \mathbb{R}^{3\times W\times H}$ is fisheye distort image captured by camera $\mathcal{C}_j$ at time t, ConvLSTM indicates single-layer ConvLSTM, PDec indicates pose decoder, $\hat{\mathbf{R}}_{ji}^{(t)}$ is the estimation of spatial rotation from $\mathcal{C}_i$ to $\mathcal{C}_j$ at time t, T is the length of image sequences.

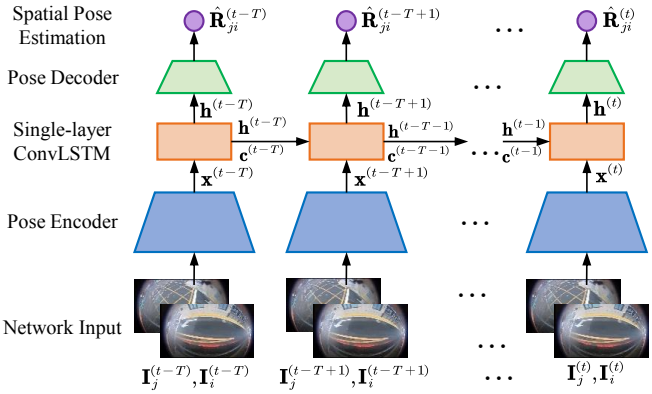

Fig. 4. ArticuNet network architecture: pose encoder takes two distort images as input, pose decoder outputs the spatial rotation transformation between cameras capturing input images.

As six fisheye cameras compose four pairs of cameras in aim of two articulation angle estimations, the proposed network acts as a siamese network that is weight-sharing among four pairs of images. In addition, as spatial translation can be obtained from spatial rotation transformation given in (4), network thus only outputs spatial rotation.

C. Loss Function

As previously deducted spatial pose transformation $\mathbf{R}_{ji}$ and $\mathbf{t}_{ji}$ both imply articulation transformation ${}^{y}\mathbf{R}_{ji}(\theta)$, we thus implement supervision over both variables. The loss function can be written as follows,

$$\mathcal{L}_{ji}^{(t)}(\theta) = \|\hat{\mathbf{R}}_{ji}^{T}\mathbf{R}_{ji} - \mathbf{I}\|_{\texttt{Frob}} + \lambda\|\hat{\mathbf{t}}_{ji} - \mathbf{t}_{ji}\|_{\mu}, \tag{8}$$

where $\hat{\mathbf{R}}_{ji}$ and $\mathbf{R}_{ji}$ indicates estimated and groundtruth spatial rotation transformation, respectively, Frob implies Frobenius norm, $\mathbf{I}$ is identity matrix, γ is distance norm in Euclidean space, $\hat{\mathbf{t}}_{ji}$ implies estimated spatial translation

and λ is the weighting coefficient balancing two losses. For simplicity, we lose super- and subscript for some notions.

The estimated spatial translation $\hat{\mathbf{t}}_{ji}$ can be obtained by substituting groundtruth $\mathbf{R}_{ji}(\theta)$ with the estimated $\hat{\mathbf{R}}_{ji}(\hat{\theta})$ in (4), where the estimation can be derived with estimated spatial rotation and by revising (1) as follows,

$$ {}^{y}\hat{\mathbf{R}}_{ji}(\hat{\theta}) = {}^{x}\mathbf{R}_{j}(\gamma) \cdot \hat{\mathbf{R}}_{ji} \cdot {}^{x}\mathbf{R}_{i}^{T}(\gamma), \tag{9} $$

$$ \hat{\mathbf{t}}_{ji} = {}^{x}\mathbf{R}_{j}^{T}(\gamma) \cdot \left[{}^{y}\hat{\mathbf{R}}_{ji}(\hat{\theta}) \cdot \left(\frac{\mathbf{t}_w}{2} - \mathbf{t}_l \right) - \left(\frac{\mathbf{t}_w}{2} + \mathbf{t}_l \right) \right]. \tag{10} $$

Since each articulation angle is formed by two adjacent carbodies, accompanied by two pairs of spatially adjacent fisheye cameras, the overall loss is averaged over the number of camera pairs aimed for the estimations, and the length of image sequences as follows,

$$ \mathcal{L} = \frac{1}{T} \sum_{t} \sum_{ji} \mathcal{L}_{ji}^{(t)}(\theta), \tag{11} $$

where $\theta \in \{\theta_f, \theta_r\}$. $ji \in \{\mathcal{C}_4\&\mathcal{C}_2, \mathcal{C}_1\&\mathcal{C}_3\}$ when $\theta = \theta_f$; $ji \in \{\mathcal{C}_6\&\mathcal{C}_4, \mathcal{C}_3\&\mathcal{C}_5\}$ when $\theta = \theta_r$.

D. Data Imbalance Solution

To deal with the imbalance between sharp turnings (large articulation angle) and mostly straight traveling cases (small articulation angle), we leveraged Gaussian-fitted probability density function [17] (see Fig. 5) of articulation angles in trainset to penalize more on loss whose groundtruth articulation angles are sparsely distributed. The overall loss can now be rewritten as follows,

$$ \mathcal{L} = \frac{1}{T} \sum_{t} \sum_{ji} (1 - \mathcal{N}(\theta)) \cdot \mathcal{L}_{ji}^{(t)}(\theta), \tag{12} $$

where $\mathcal{N}(\theta) \sim (\mu, \sigma^2)$ is the Gaussian-fitted PDF, μ and σ^2 implies sample mean and sample variance, respectively.

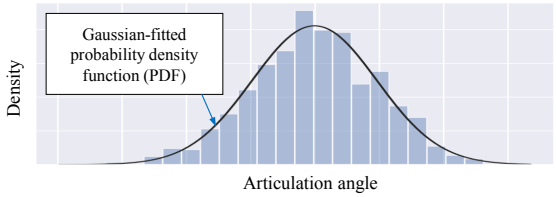

Fig. 5. Using Gaussian distribution to fit the probability density function of observed articulation angles in trainset. Blue bars show the histogram, black curve indicates fitted Gaussian distribution.

E. Articulation Angle Recovering

As the network outputs estimated spatial pose transformation in the pattern of a rotation matrix ${}^{y}\hat{\mathbf{R}}_{ji}(\hat{\theta})$, to recover articulation angle, decomposition from rotation matrix to Euler angle should be implemented [18]. Though Euler angles suffer from Gimbal Lock problem at $\pm 90°$ [19], this does not affect the validity of our proposal, as the articulation angles are normally ranged from $-36°$ to $36°$ due to physical constraint by real-time operation. Therefore, the general 3×3 rotation matrix can be decomposed into a combination of sequential rotations around each axis: roll (ϕ), yaw (θ), and pitch (ψ), which could be summarised as $\mathbf{R} = {}^{z}\mathbf{R}(\phi) \cdot {}^{y}\mathbf{R}(\theta) \cdot {}^{x}\mathbf{R}(\psi)$. By decomposing ${}^{y}\hat{\mathbf{R}}_{ji}(\hat{\theta})$, roll and pitch angles are expected to be zero, as the vehicle is presumed to carry out 2D planar motion, making $\hat{\theta}$ our ultimate estimation objective. Furthermore, as we previously labeled different camera pairs on the same carbody with different signs of groundtruth articulation relationship, by converting to final estimations, inverse signed operation should be implemented when necessary, as in, $\hat{\theta} \leftarrow -\hat{\theta}$.

IV. Experiments

A. Dataset

Dataset used for supervised articulation angle estimation was collected based on the panoramic camera system of Shanghai Lingang DRT, a three-carbody multi-articulated rubber tyred low floor Digital Railway Tram (DRT) manufactured by CRRC Nanjing Puzhen which runs on Shanghai Lingang T2 Line (see Fig. 6). Time-synchronized 6 tracks of unrectified fisheye images as well as values of two articulation angles were recorded on the whole coverage of T2 Line and at the frequency of 15 Hz. Data consist of 33909 frames and were utterly divided into trainset, validation set, and testset in proportion of 80%, 10%, and 10%, respectively.

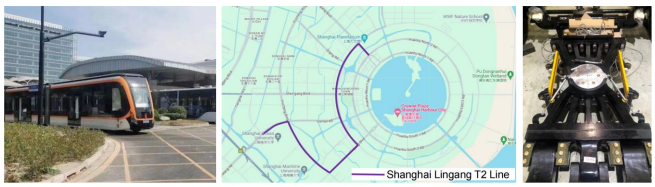

Fig. 6. (a) Shanghai Lingang Digital Rail Train operating on T2 Line; (b) Shanghai Lingang T2 Line route map; (c) Mechanical angle sensor.

Distribution of observed articulation angles and Gaussian-fitted PDF are shown in Fig. 7, where large articulation angles could be found to be sparsely distributed.

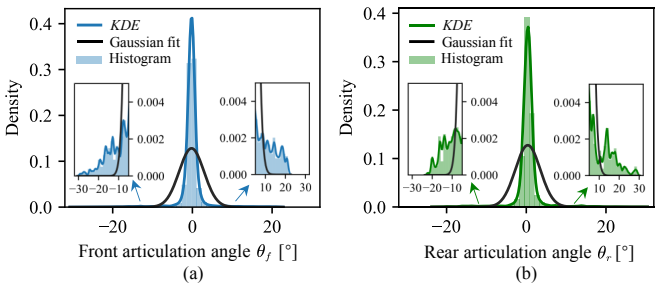

Fig. 7. Distribution of (a) front articulation angle θ_f; (b) rear articulation angle θ_r in training set. Black curve indicates fitted Gaussian PDF.

B. Training Settings

Evaluation Metrics. Mean Absolute Error (MAE), RMSE (Root Mean Squared Error) and Mean Absolute Percentage Error (MAPE) of articulation angle θ, spatial rotation error

TABLE I

EXPERIMENT RESULTS ON TESTSET (# DENOTES USING PRETRAINED POSE ENCODER ON IMAGENET).

Angle	Pair	θ_{MAE} [°] ↓	θ_{MAPE} [%]	θ_{RMSE} [°] ↓	$\epsilon_{\mathbf{R}}$ ↓	$\epsilon_{\mathbf{t}}$ ↓	$\max \epsilon_\theta$ [°] ↓	$\max \epsilon_\psi$ [°] ↓	$\max \epsilon_\phi$ [°] ↓
$\theta_f^{\#}$	$\mathcal{C}_4$&$\mathcal{C}_2$	0.136	0.649	0.228	3.358	0.011	2.934	0.054	0.024
	$\mathcal{C}_1$&$\mathcal{C}_3$	0.134	0.645	0.232	3.306	0.011	2.424	0.067	0.019
	average	0.112	0.544	0.187	N/A	N/A	2.145	0.009	0.013
θ_f	$\mathcal{C}_4$&$\mathcal{C}_2$	0.155	0.672	0.274	3.227	0.013	2.454	0.020	0.011
	$\mathcal{C}_1$&$\mathcal{C}_3$	0.170	0.668	0.338	3.269	0.014	4.415	0.022	0.006
	average	0.133	0.550	0.241	N/A	N/A	2.202	0.003	0.008
$\theta_r^{\#}$	$\mathcal{C}_6$&$\mathcal{C}_4$	0.128	0.415	0.230	3.606	0.010	2.563	0.061	0.025
	$\mathcal{C}_3$&$\mathcal{C}_5$	0.129	0.409	0.228	3.537	0.011	2.838	0.080	0.021
	average	0.108	0.357	0.190	N/A	N/A	2.009	0.067	0.001
θ_r	$\mathcal{C}_6$&$\mathcal{C}_4$	0.156	0.460	0.318	3.455	0.013	4.782	0.021	0.010
	$\mathcal{C}_3$&$\mathcal{C}_5$	0.162	0.525	0.330	3.459	0.013	4.703	0.023	0.009
	average	0.132	0.401	0.249	N/A	N/A	3.255	0.002	0.008

$\epsilon_{\mathbf{R}}$, spatial translation error $\epsilon_{\mathbf{t}}$ as well as maximum estimation errors of articulation angle, pitch angle and roll angle are used as evaluation metrics. Maximum estimation errors were included to illustrate how extreme cases have been handled by proposed algorithm. Calculation rules are as follows,

$$\theta_{\text{MAE}} = \frac{1}{M}\sum_{m=1}^{M}|\theta_m - \hat{\theta}_m|, \tag{13}$$

$$\theta_{\text{MAPE}} = \frac{100}{M}\sum_{m=1}^{M}\left|\frac{\theta_m - \hat{\theta}_m}{\theta_m}\right|, \tag{14}$$

$$\theta_{\text{RMSE}} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\theta_m - \hat{\theta}_m\right)^2}, \tag{15}$$

$$\epsilon_{\mathbf{R}} = \|\hat{\mathbf{R}}_{ji}^T\mathbf{R}_{ji} - \mathbf{I}\|_{\texttt{Frob}}, \tag{16}$$

$$\epsilon_{\mathbf{t}} = \|\hat{\mathbf{t}}_{ji} - \mathbf{t}_{ji}\|_2, \tag{17}$$

$$\max \epsilon_\alpha = \max_m |\alpha_m - \hat{\alpha}_m|,\ \alpha \in \{\theta, \psi, \phi\}, \tag{18}$$

where θ_m and $\hat{\theta}_m$ are groundtruth and estimated articulation angle, respectively, M is the number of data.

Implementation Details. We implement ArticuNet in PyTorch. A on ImageNet [20] pretrained ResNet18 backbone [21] was used as pose encoder. Pose decoder consists of three stacked-up 2D convolutions. ArticuNet was trained on a single NVIDIA Tesla P100 GPU with Adam optimizer for 10 epochs. Learning rate was set to 1×10^{-4} with exponential learning rate decay where the decay coefficient equals 0.1. Input unrectified images are in resolution of 960×540, where they would be firstly resized to 480×270 to reduce computation cost. Batch size is set to 8. λ is set to 0.9 to balance rotation loss and translation loss. Distance norm μ is set to 2. Installation angle γ is 17°. Physical scale-to-meter distances $s_l = 4.5$ and $s_w = 2.5$. Length of history image sequences is set $T = 6$. Color adjustment has been used as data augmentation during training.

C. Experiment Results

Table I lists the experiment results of ArticuNet on testset, where comparisons of metrics evaluating the estimations regarding both front and rear articulation angles have been made between the case of using pretrained pose encoder and that trained from scratch. As one angle has two camera pairs available for estimations, we list out the estimation accuracy of each individual camera pair as well as their average case. Spatial rotation and translation error, denoted as $\epsilon_{\mathbf{R}}$ and $\epsilon_{\mathbf{t}}$, respectively, would not be subjected to averaging, given that the rotation and translation transformation varies between different camera pairs. Results listed in Table I indicate accurate estimations of our proposed network especially with pose encoder pretrained on ImageNet. Model's estimation accuracy could also be attributed to the fact that fisheye cameras primarily capture near-field images of the vehicle, this, however, could result in less diversity in feature distributions compared to images of strictly front or side view cameras, where more traffic agents are involved. In addition, as each articulation angle has two camera pairs available for estimations, averaging their outputs could be observed to contribute to fewer deviations from the true values. By practical implementation and application, this setting could be feasible and beneficial in greater robustness against estimation errors, where the estimation outcomes could be programmed to refer to one another and meanwhile also compensate for estimation inaccuracy and imprecision.

Furthermore, as the direct output of the network is in the form of a rotation matrix, by decomposing it into Euler angles, pitch and roll angles could also be obtained. As the vehicle is presumed to perform 2D planar motions, pitch and roll angles are presumptuously labeled as zero. To verify whether the network was trained to learn pure rotation around the y-axis (of the camera), maximum absolute estimation errors of both front and rear pitch angle ψ and roll angle ϕ were thus calculated and compared. Results listed in Table I shows estimations of almost pure rotation around the y-axis, as maximum absolute estimation errors of pitch and roll angles are all observed to be significantly close to zero. In addition, results of $\max \epsilon_\theta$ show the estimation performance in handling extreme cases, contrasting with metrics averaged

over the entire test set, such as θ_{MAE}. This contrast is due to the imbalanced distribution of data, further demonstrating the deprived learning on large articulation angle cases compared to small articulation angle cases. In following ablation study, we demonstrated how the proposed distribution-weighted loss function alleviated maximum estimation error.

D. Field Experiment

In Fig. 8, we present the estimated results for both front and rear articulation angles over a 60-second long sequence where the articulated vehicle was executing sharp-angle turns from a field experiment beyond the collected dataset. The estimations were generated by averaging the estimations obtained from both pairs of cameras utilized for estimations. The results demonstrated small deviations from the groundtruth, as recorded by angle sensors.

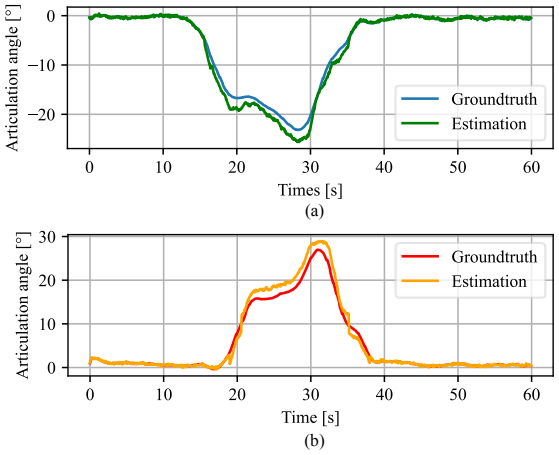

Fig. 8. estimations of articulation angles from a field experiment: (a) front articulation angle θ_f; (b) rear articulation angle θ_r.

E. Ablation Studies

Data imabalance. Here we substitute Gaussian-fitted PDF with PDF of a uniform distribution, indicating each angle observed in trainset has the same probability density, to find out how distribution balancing could influence results on testset. PDF of uniform distribution $\theta \sim \mathcal{U}(\min_\theta, \max_\theta)$ is given as $\mathcal{U}(\theta) = 1/(\max_\theta - \min_\theta)$. Thus, (12) should be rewritten as follows,

$$\mathcal{L} = \frac{1}{T} \sum_t \sum_{ji} (1 - \mathcal{U}(\theta)) \cdot \mathcal{L}_{ji}^{(t)}(\theta), \tag{19}$$

where $\min_\theta$ and $\max_\theta$ are the minimum and maximum of observed articulation angles in trainset.

Similarly, we use maximum absolute estimation error for ablation analysis. Abalation results presented in Fig. 9 show increases in maximum absolute estimation error $\max \epsilon_{\theta_f}$ and $\max \epsilon_{\theta_r}$ when treating each groundtruth equivalently in overall distribution, verifying benefit of using Gaussian-fitted PDF to weigh and penalize on losses whose groundtruth articulation angles are sparsely distributed.

Temporal dependency. A comparison of estimation performance was conducted between models with and without

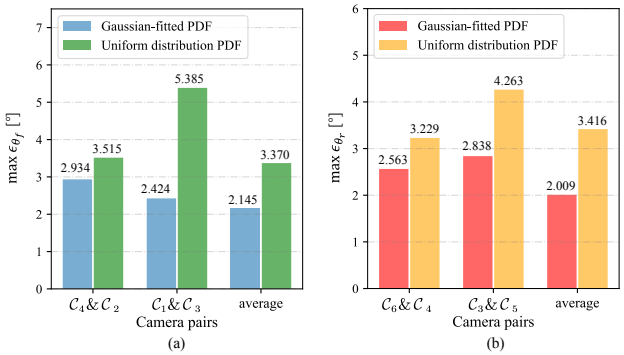

Fig. 9. Result of ablation study on data imbalance: maximum estimation error of (a) front articulation angle θ_f and (b) rear articulation angle θ_r.

TABLE II

ABLATION STUDY ON TEMPORAL DEPENDENCY (W AND W/O INDICATES WITH AND WITHOUT CONVLSTM FOR TEMPORAL DEPENDENCY)

Angle	Method	θ_{MAE} [°] ↓	θ_{RMSE} [°] ↓	$\max \epsilon_\theta$ [°] ↓
$\theta_f^{\#}$	w. T	0.112	0.187	2.145
	w/o. T	0.176	0.297	3.785
$\theta_r^{\#}$	w. T	0.108	0.190	2.009
	w/o. T	0.162	0.288	3.866

ConvLSTM for temporal dependency establishment. By removing the ConvLSTM layer, ArticuNet estimates the spatial pose independently for each frame. The results listed in Table II demonstrate the network's capability of estimating the articulation angle based solely on images, but with degraded estimation accuracy and performance compared to the proposed network design in handling extreme cases and smoothing out abrupt changes in estimations.

Spatial rotation representation. Quaternion can also be utilized to represent spatial rotation transformation. In this case, we also made an attempt of Quaternion to realize supervised training. Network architecture was altered to align the output dimension with that of Quaternion. Here we denote Quaternion concerning installation angle γ of $\mathcal{C}_i$ as ${}^x\mathbf{q}_i(\gamma) = [\cos(\pi/4 - \gamma/2), \sin(\pi/4 - \gamma/2), 0, 0]^T$ and the articulation relationship between adjacent carbodies as ${}^y\mathbf{q}_{ji}(\theta) = [\cos(\theta/2), 0, \sin(\theta/2), 0]^T$. Therefore, (1) and loss function (8) can be rewritten as follows in (20) and (21), respectively. In addition, as a single rotation can be mapped from two Quaternions ($\mathbf{q}$ and $-\mathbf{q}$, each from one hemisphere) [22], we confined all Quaternions to one hemisphere to ensure a unique value for each rotation.

$$\mathbf{q}_{ji} = {}^x\mathbf{q}_j^*(\gamma)\ {}^y\mathbf{q}_{ji}(\theta)\ {}^x\mathbf{q}_i(\gamma) \tag{20}$$

$$\mathcal{L}_{ji}^{(t)}(\theta) = \|\mathbf{q}_{ji} - \hat{\mathbf{q}}_{ji}\|_2 \tag{21}$$

Output of the network is denoted as ${}^y\hat{\mathbf{q}}_{ji}$, and the estimated articulation relationship can be derived as follows in (22). Furthermore, articulation angle, pitch and roll angles could be recovered by decomposing the estimation ${}^y\hat{\mathbf{q}}_{ji}(\hat{\theta})$,

$${}^y\hat{\mathbf{q}}_{ji}(\hat{\theta}) = {}^x\mathbf{q}_j(\gamma)\ \hat{\mathbf{q}}_{ji}\ {}^x\mathbf{q}_i^*(\gamma). \tag{22}$$

TABLE III

EXPERIMENTAL RESULTS ON THE TESTSET USING QUATERNIONS AS SPATIAL ROTATION REPRESENTATION.

Angle	Pair	θ_{MAE} [°] ↓	θ_{MAPE} [%]	θ_{RMSE} [°] ↓	$\epsilon_{\mathbf{q}}$ ↓	$\max \epsilon_\theta$ [°] ↓	$\max \epsilon_\psi$ [°] ↓	$\max \epsilon_\phi$ [°] ↓
$\theta_f^{\#}$	$\mathcal{C}_4$&$\mathcal{C}_2$	0.148	0.654	0.253	0.001	3.178	0.111	0.107
	$\mathcal{C}_1$&$\mathcal{C}_3$	0.155	0.749	0.268	0.001	4.136	0.100	0.151
	average	0.129	0.581	0.228	N/A	3.657	0.088	0.117
$\theta_r^{\#}$	$\mathcal{C}_6$&$\mathcal{C}_4$	0.158	0.476	0.283	0.002	2.827	0.102	0.186
	$\mathcal{C}_3$&$\mathcal{C}_5$	0.160	0.509	0.272	0.002	3.817	0.125	0.169
	average	0.125	0.385	0.230	N/A	2.805	0.097	0.173

In addition to MAE, MAPE and RMSE, spatial rotation error was evaluated using estimated and groundtruth Quaternions. The calculation rule is given as follows in (23),

$$\epsilon_{\mathbf{q}} = \left\| \frac{1}{\|\hat{\mathbf{q}}_{ji}\|\|\mathbf{q}_{ji}\|} \hat{\mathbf{q}}_{ji}^{*} \ \mathbf{q}_{ji} - \mathbf{q}_e \right\|_2, \tag{23}$$

where $\mathbf{q}_e = [1, 0, 0, 0]^T$.

Results of using Quaternion for supervised learning are listed in Table III, which exhibits fairly similar estimation accuracy with our proposed method using rotation matrix. This confirms the effectiveness of using other representations of spatial pose transformation to realize estimations.

V. CONCLUSION

This study proposed a deep-learning-based articulation angle estimation solution for multi-articulated vehicles based on panoramic camera system. The proposed network ArticuNet takes images captured by spatially adjacent cameras as inputs, uses ConvLSTM to establish temporal dependency and outputs the spatial rotation transformation matrix, where articulation angles could be subsequently obtained by rotation matrix decomposition to Euler angles. Datasets used for network training were recorded on T2 Line of Shanghai Lingang DRT. A Gaussian-fitted probability density function of the articulation angles in trainset was adopted to weigh and penalize the loss function in correspondence to the distribution density of groundtruth articulation angle. The experiment's results demonstrated precise and accurate articulation angle estimation, providing further verification for the viability of employing a vision-based approach as a substitute for mechanical angle sensors.

REFERENCES

[1] K. Olutomilayo and D. R. Fuhrmann, "Estimation of trailer-vehicle articulation angle using 2d point-cloud data," in *2019 IEEE Radar Conference (RadarConf)*. IEEE, 2019, pp. 1–6.

[2] Z. Sun, P. Dai, Z. Tian, H. Meng, and W. Wang, "Steering modeling and off-tracking suppression control of train-like automobiles," *IFAC-PapersOnLine*, vol. 55, no. 37, pp. 223–228, 2022.

[3] J. Feng and Z. Sun, "Path-tracking control and following control of tractor-semitrailer combination based on improved mpc," SAE Technical Paper, Tech. Rep., 2023.

[4] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[5] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[6] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia, "Monovit: Self-supervised monocular depth estimation with a vision transformer," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 668–678.

[7] X. Peng, J. Cui, and L. Kneip, "Articulated multi-perspective cameras and their application to truck motion estimation," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 2052–2059.

[8] 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.

[9] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, Y. Rao, G. Huang, J. Lu, and J. Zhou, "Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation," in *Conference on Robot Learning*. PMLR, 2023, pp. 539–549.

[10] K. T. Olutomilayo, M. Bahramgiri, S. Nooshabadi, J. Oh, M. Lakehal-Ayat, D. Rogan, and D. R. Fuhrmann, "Trailer angle estimation using radar point clouds," *Signal Processing*, vol. 188, p. 108221, 2021.

[11] C. de Saxe and D. Cebon, "Camera-based articulation angle sensing for heavy goods vehicles," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 8, pp. 7522–7535, 2021.

[12] C. Fuchs, S. Eggert, B. Knopp, and D. Zöbel, "Pose detection in truck and trailer combinations for advanced driver assistance systems," in *2014 IEEE Intelligent Vehicles Symposium Proceedings*. IEEE, 2014, pp. 1175–1180.

[13] C. Fuchs, D. Zöbel, and D. Paulus, "3d pose detection for articulated vehicles," in *Intelligent Autonomous Systems 13: Proceedings of the 13th International Conference IAS-13*. Springer, 2016, pp. 459–472.

[14] Y. Atoum, J. Roth, M. Bliss, W. Zhang, and X. Liu, "Monocular video-based trailer coupler detection using multiplexer convolutional neural network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5477–5485.

[15] L. Kaci, B. Samet, M. Yahiaoui, A. Makni, and H. D. Mousselmal, "Image based vehicle-trailer angle estimation," in *2022 IEEE 9th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*. IEEE, 2022, pp. 286–290.

[16] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[17] M. I. Ribeiro, "Gaussian probability density functions: Properties and error characterization," *Institute for Systems and Robotics, Lisboa, Portugal*, 2004.

[18] G. G. Slabaugh, "Computing euler angles from a rotation matrix," *Retrieved on August*, vol. 6, no. 2000, pp. 39–63, 1999.

[19] S. L. Altmann, *Rotations, quaternions, and double groups*. Courier Corporation, 2005.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5974–5983.