



SLAM Visual Localization and Location Recognition Technology Based on 6G Network

Cheng Zhang¹ · Yuchan Yang² · Guangyao Li¹

Accepted: 10 May 2024

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

Sixth-generation (6G) wireless systems, when ultimately deployed, will comprise intelligent wireless networks that provide high-accuracy localization services together with ubiquitous communication. By bringing in a fresh set of traits and functionalities that allow location and communication to coexist while sharing resources, they provide the impetus for this change. By identifying the critical technological enablers that open up exciting new possibilities for combined localization and sensing applications, we concentrate on converged 6G communication, localisation, and sensing systems. 6G will advance toward even higher frequency ranges, broader bandwidths, and massive antenna arrays in terms of potential enabling technologies. Owing to the drawbacks of LiDAR, including its high price, short lifespan, and large volume, visual sensors—inexpensive and lightweight—are garnering increased interest and developing into a hotspot for study. With the rapid advancements in deep learning (DL) and hardware computing capacity, new approaches and concepts for solving visual simultaneous localization and mapping (VSLAM) difficulties have surfaced. We concentrate on the visual odometry (VO) application of DL and VSLAM integration. Most VO algorithms used today, such as those for motion estimation, feature extraction, feature matching, local optimization, etc., are created using subpar pipelines. Using Convolution LSTM, a unique end-to-end design for monocular VO is presented in this research. It does not adopt any module in the traditional VO pipeline, instead inferring postures directly from a series of raw RGB photos (videos) because it has been trained and deployed end-to-end. It uses CNN to automatically train an adequate representation of features for the VO problem based on the Convolution LSTM, which is utilized to simulate sequential dynamics and relations implicitly. Comprehensive tests on the KITTI VO dataset demonstrate competitive performance compared to cutting-edge techniques. This confirms that the end-to-end DL approach can be a viable addition to conventional VO systems.

Keywords 6G · SLAM · Visual SLAM · Deep learning · Visual odometry · CNN and AlexNet

✉ Guangyao Li
Lligyao@163.com

¹ College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China

² Lotus Robotics, Hangzhou 310051, Zhejiang, China

1 Introduction

Robots navigating through an unfamiliar environment can accomplish self-localization and mapping thanks to a technology known as simultaneous localization and mapping (SLAM). Sonar and LiDAR sensors, which offer a high degree of accuracy but are heavy, expensive, and fragile, were a major component of SLAM in its early stages. On the other hand, visual sensors became an alternative since they are small, inexpensive, and simple to use. To facilitate location and navigation in challenging real-world contexts, visual simultaneous localization and mapping, or VSLAM, can use visual sensors that function similarly to human eyes to sense the surrounding environment and gather rich environmental information. VSLAM technology is important for several applications, such as military rovers, drones, unmanned vessels, intelligent robotics, autonomous cars, augmented reality (AR), and virtual reality (VR) [1]. Furthermore, the physical world and virtual cyberspace are interwoven thanks to recent advancements in AR and VR technology. In addition to maintaining the overlay virtual items' geometric coherence with the physical world, the 3D map rebuilt by VSLAM can include geometric details regarding the scene, enhancing the realism of the virtual environment. VSLAM technology is becoming increasingly in demand, driving the emergence of new techniques and technologies and making it a popular topic for research.

Throughout the past few decades, there has been a great interest in both the computer vision and robotics communities in visual odometry (VO), one of the most important approaches for posture estimation and robot localization [2]. It has been widely used as an addition to GPS, Inertial Navigation System (INS), wheel odometry, and other systems on various robots. Wireless networks are often lauded for their communication capabilities only, ignoring their innate benefits related to localization and sensing. With its enormous antenna array, high carrier frequency, and big bandwidth, the 5G NR access interface presents excellent prospects for precise localization and sensing systems in this area. Furthermore, 6G systems will carry on the trend of operating at increasingly higher frequencies, such as those at the millimeter wave (mmWave) and THz1 bands, and with even bigger bandwidths. The THz frequency range presents excellent prospects for frequency spectroscopy, high-definition imaging, and precise localization. The authors of [3] summarise wireless communications and the intended uses for 6G networks that operate above 100 GHz. They then discuss the potential of mmWave and THz-enabled localization and sensing solutions. Similarly, [4] discusses potential paths the cellular industry may take to develop future 6G systems.

With the emergence of 6G frameworks, the line between communication from localization is becoming more and more blurred, necessitating the development of seamless integration solutions. This convergence offers improved usefulness and efficiency, enabling a wide range of applications ranging from augmented reality to driverless vehicles. Although very efficient, current localization technologies like LiDAR are constrained by a number of issues, including expensive operating expenses, large physical bulk, and comparatively limited operational lifetimes. Their scalability and general adoption are limited by these disadvantages, especially in consumer-grade applications. Because of their affordability, portability, and adaptability, visual sensors offer a viable substitute. Their potential has been greatly increased by developments in camera and image processing, which makes them perfect for integration into mobile and ubiquitous computing environments. This research finds solution for,

- How can the integration of deep learning with visual odometry be optimized to take full advantage of 6G capabilities?
- What are the specific advantages and challenges of using visual sensors over traditional localization technologies like LiDAR in the context of 6G?
- Can an end-to-end deep learning model effectively replace traditional multi-stage VO systems without compromising accuracy and reliability?

Deep Learning (DL) has shown encouraging results dominating numerous computer vision tasks. Unfortunately, this hasn't arrived yet for the VO problem. Not much work has been done on VO—not even about 3D geometry issues. This is likely because most models that have been trained and DL architectures currently in use are primarily built to address recognition and classification tasks, which motivates deep convolutional neural networks (CNNs) for extracting high-level visual data from images. Understanding appearance representation severely impedes the VO's ability to become widely known and restricts its use in controlled situations. For this reason, the VO algorithms mostly rely on geometric properties rather than visual ones. Instead of processing a single image, an AV algorithm should ideally describe motion dynamics by looking at the changes and linkages in a sequence of images. This implies that sequential learning is required, which the CNNs cannot provide. To satisfy these needs, this paper contributes the following:

- This paper proposed a novel end-to-end monocular VO approach using convolution LSTM in 6G wireless communication systems.
- This study uses DL approaches to demonstrate the monocular VO problem in an end-to-end way (directly determining the poses from the RGB images).
- The input from the captured video sequence or RGB image is preprocessed to remove the noise. Then, the new geometric features from the RGB images are mapped and extracted using Global channel attention and CNN methodologies.
- Long-short-term memory (LSTM) intuitively captures and automatically understands the sequential dependencies and complicated motion dynamics of an image series, which are important to the VO but cannot be openly or simply modelled by humans.
- The developed model is experimented with using the KITTI dataset, and the model's efficiency is discussed.

The remainder of the paper is structured as follows: Sect. 2 reviews related work. Section 3 describes the end-to-end monocular-VO method with preprocessing, feature mapping and extraction, and sequence modelling. Section 4 presents the experimental findings. Section 5 concludes.

2 Related Work

Author [5] provide a comprehensive overview of the various SLAM technologies implemented for AV perception and localisation. The authors also offer a comprehensive review of various V SLAM schemes, their strengths and weaknesses, as well as the challenges of deployment of V SLAM and future research directions. Author [6] highlight important technological enablers for convergent 6G communication, localisation, and sensing systems, review their underlying difficulties and implementation concerns, and suggest possible solutions. We also review the fascinating new prospects for integrated localisation and

sensing applications, which will upend conventional design ideas and fundamentally alter how we live, interact with our surroundings, and conduct business. 6G will advance toward even higher frequency ranges, broader bandwidths, and huge antenna arrays in terms of potential enabling technologies. Consequently, this will allow for sensing systems with high Doppler, angle, and range resolutions and accurate localisation down to the centimetre level.

Author [7] examine the potential uses and applications of localization in upcoming 6G wireless systems and explore the effects of the key technological enablers. Next, system models considering line-of-sight (LOS) and non-LOS channels are offered for millimetre wave, terahertz, and visible light placement. Additionally, mathematical definitions and a review of localization key performance indicators are provided. A thorough analysis of the most advanced conventional and learning-based localisation approaches is also carried out. In addition, the design of the wireless system is taken into account, the localisation problem is stated, and their optimisation is looked at Author [1]. This research thoroughly analyses deep learning-based VSLAM techniques. We describe the basic ideas and framework of VSLAM and briefly overview its development process. Next, we concentrate on the three parts of deep learning and VSLAM integration: mapping, loop closure detection, and visual odometry (VO). We provide a detailed summary and analysis of each algorithm's strengths and weaknesses. Furthermore, we offer an overview of commonly utilised datasets and assessment metrics. Lastly, we review the unsolved issues and potential paths for merging deep learning and VSLAM.

Author [8] initially provide a detailed overview of the research findings on the subject of visual SLAM, divided into three categories: deep learning-enhanced SLAM, dynamic SLAM, and static SLAM. To sort out the fundamental technologies related to the use of 5G ultra-dense system to offload complex computing tasks from visual SLAM systems to edge computing servers, the second section of the technology contrast between mobile edge computing and mobile cloud computing, along with the sections on 5G ultra-dense networking technology and MEC and UDN integration technology, are introduced. Author [9] present OTE-SLAM, an object-tracking augmented visual SLAM system that follows dynamic objects' movements and the camera's motion. Moreover, we jointly optimise the 3D position of the item and the camera posture, allowing object tracking and visual SLAM to work together to both benefits. Experiment findings show that the suggested method enhances the SLAM system's accuracy in difficult dynamic situations.

The Extended Kalman Filter (EKF) is a valuable tool, especially when tackling nonlinear systems, as it linearizes them around the current estimate [10, 11]. Multisensor integrated navigation refers to the fusion of data from multiple sensors to determine the position, orientation, or trajectory of a vehicle or device [12, 13]. This process often involves specific metrics or measures to evaluate the effectiveness of privacy-preserving techniques [14, 15]. Accurate passenger counting holds significance across various applications like public transportation, ride-sharing services, and traffic management [16, 17].

Urban heat prediction is crucial for understanding and mitigating the effects of heat islands, areas with significantly higher temperatures due to human activities and infrastructure [18, 19]. Light field image depth estimation tasks involve estimating depth information captured in a scene [20, 21], particularly essential for applications like 3D reconstruction, autonomous driving, and augmented reality, where precise depth information is pivotal [22, 23].

Transformers represent a specific architecture widely used for sequence modeling tasks such as natural language processing or image recognition [24, 25]. Detecting glass surfaces finds utility in diverse applications such as robotics, augmented reality, or

autonomous driving, where accurate scene understanding is indispensable [26, 27]. IoT environments encompass various applications, including smart homes, industrial automation, healthcare, and smart cities [28, 29]. Image feature extraction plays a pivotal role in many computer vision tasks like object detection, classification, and segmentation [30, 31].

Adapting a traffic object detection model from one domain to another typically involves gradually refining the adaptation process from coarse adjustments to fine-tuned adjustments [32, 33]. A maximum reduction of 22% and 33% in absolute and relative trajectory inaccuracy is one of the enhancements.

3 Methodologies

This section provides a detailed description of the deep RCNN framework that realises the monocular VO in an end-to-end manner. It is mainly made up of CNN-based feature extraction, GC-based feature mapping, and LSTM-based sequential modelling. The overview of the proposed architecture flow is shown in Fig. 1. The input monocular image sequence from the video clip is taken as input. Next, the input image is pre-processed to remove noise, resize, and smooth it in the preprocessing stage. Then, the feature map is determined using the Global-CA method from the preprocessed image. The features from the feature map are extracted using CNN, and the model sequence learning is performed using LSTM. Each image pair estimates the pose at each time step t through the network. The process is repeated at each time step $t + 1$ and new poses are estimated.

Hard challenges are overcome by,

- Terahertz (THz) frequency ranges are anticipated to be used by 6G, which may enable more accurate localization with possibly centimeter-level precision. Massive MIMO (Multiple Input Multiple Output) technology, which can improve the capacity and dependability of wireless communications and location, is also made easier by these higher frequencies.
- In contrast to earlier generations, 6G seeks to lower latency and improve the effectiveness of both services by combining communication and location into a single framework.

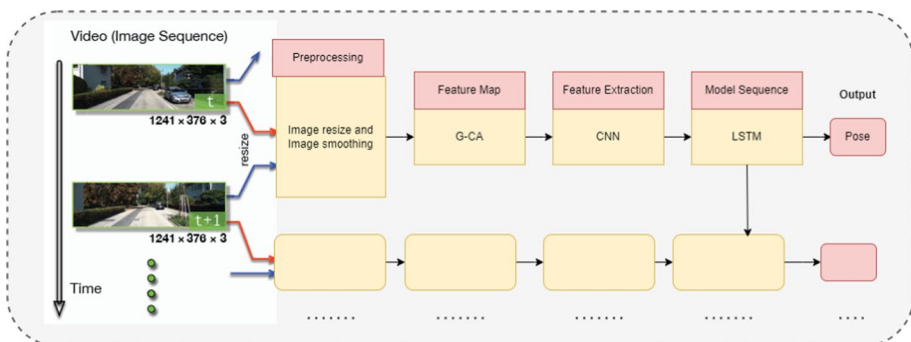


Fig. 1 Overview of the proposed VO modelling system

- Signals can be directed more precisely via beamforming, which enhances localization accuracy and lowers interference. This is especially useful in densely populated urban areas.
- With 6G, artificial intelligence (AI) and machine learning are predicted to play major roles in allowing the network to dynamically adapt to the surroundings and user needs.

3.1 Preprocessing

The input RGB image is preprocessed by subtracting the mean value of RGB values and resizing it with the new size as the multiple of 64. The partial volume effect is caused by variations in real-time applications that impact the input images. The Bias field detection and correction approach is utilised to get around this. The difference between the grey pixels of comparable tissues is known as the bias field and is seen as the picture multiplicative module. Recent studies on RGB images have shown that smoothing improves results compared to non-smoothing methods. Therefore, this paper pretreated RGB images to improve feature extraction and sequencing results using the bias field reduction and smoothing procedure.

The noise N and bias B of the true images of x_0 and x_t is written as in Eq. (1)

$$x_t = Bx_0 + N \quad (1)$$

Once the bias field is identified, it is corrected using N4ITK method [34]. To smooth the image, a Gaussian filter has been used with the kernel size of 5×5 as in Eq. (2)

$$I_{smooth}(G(x, y)) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (2)$$

where σ denotes the standard deviation.

3.2 Feature Mapping Using G-CA

Based on 1-dimensional Convolution via ECANet [35], the nonlocal neural network [36] underpins the GCA process. As illustrated in Fig. 2, given $b \times h \times w$ from the backbone network as feature tensor F . To obtain the $1 \times b$ query as Q_b and the key k_b , applied the global-average pooling (GAP) with spatial dimensions followed by the 1D convolution along the kernel size of k and sigmoid activation function. The outlier product of Q_b and k_b is formed through softmax functions over the channels to comprise $b \times b$ GCA map,

$$A_b^g = \text{softmax}(k_b^T Q_b) \quad (3)$$

At the end, the attention map is V_b as $(V_b \times A_b^g)$ which is reshaped back as $b \times h \times w$ to produce the G-CA map G_b . The channel attention is denoted as in Eq. (4)

$$G_b = \sigma(\text{Fully_Connected}(\text{Max}_{pool}(X)) + \text{Fully_Connected}(\text{Average}_{pool}(X))) \quad (4)$$

3.3 End to End VO Using C-LSTM

Several well-known and potent DNNar architectures, such VGGNet [37] and GoogleNet [38], were created for computer vision applications and have demonstrated exceptional performance. Most of them are built to solve problems related to recognition, classification,

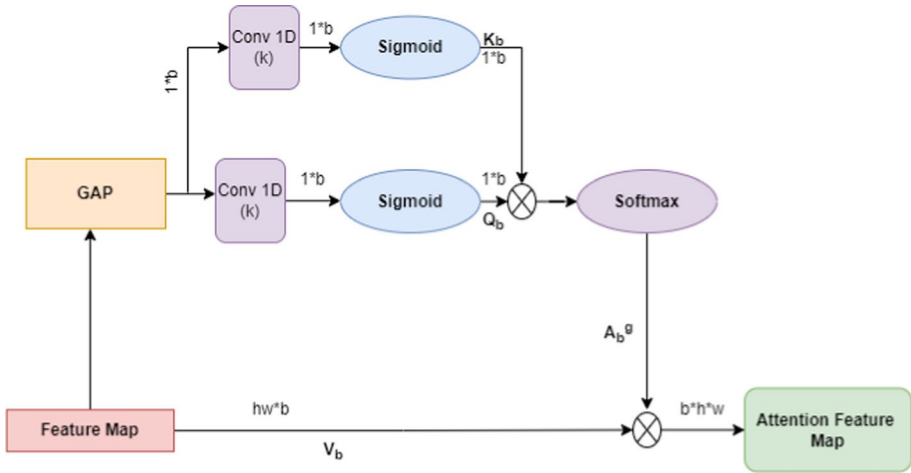


Fig. 2 G-CA

and detection; thus, they are taught to derive knowledge from appearance and visual context. However, as was previously mentioned, VO—which has its roots in geometry—should not be strongly associated with look. As such, applying the widely used DNN architectures currently available for the VO problem is not feasible. Addressing the VO and other geometric problems requires a framework to learn geometric feature representations. Nevertheless, as VO systems function on picture sequences obtained during movement, inferring relationships between successive image frames, such as motion models, is crucial. These relationships grow over time. As a result, the suggested C-LSTM takes these needs into account. The proposed end-end VO system is shown in Fig. 3.

As seen in the above diagram, the C-LSTM (Convolutional Long Short-Term Memory) architecture is a novel strategy created to address the particular difficulties associated with voice over the internet. Convolutional neural networks (CNNs) and long short-term memory (LSTM) units are used in this architecture to provide a system that can interpret spatial data and take the temporal sequence of images into account to infer motion.

- This model, specifically designed to learn geometric feature representations—which are essential for effectively simulating motion between consecutive image frames—is the C-LSTM framework. In contrast to appearance-focused architectures, C-LSTM places more emphasis on the scene’s geometry and the relative motion of the object or camera.

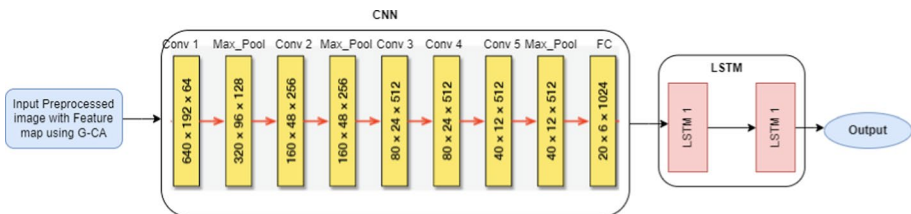


Fig. 3 Proposed end-end VO using DL

- The C-LSTM's LSTM component is especially made to record the dependencies and temporal linkages between a series of frames. This is crucial for voice over internet (VO), as trajectory estimate accuracy is strongly correlated with the comprehension of continuity and changes in position over time.
- Without the need for human feature extraction or pre-processing, the system can learn directly from raw RGB images thanks to the end-to-end training methodology. By doing this, the system may potentially become more versatile and perform better by automatically figuring out what features are most relevant for voice over internet jobs.

3.3.1 CNN (AlexNet) Based Feature Extraction

For the study of virtual images, CNN is a popular DL model [39, 40]. Generally speaking, CNN uses the image as input and divides it into many categories. Input neurons, a sequence of convolutional layers, pooling, fully connected layers, and normalising layers make up its structure [41]. The convolution layer's neurons have a tiny region connecting them to the layer before it. The fully connected layers' activation neurons are connected to the layers below. Equation (5) represents a fully connected function's forward and backwards reverse propagation.

$$x_i^{l+1} = \sum_i w_{j,i}^{l+1} x_i^l \quad (5)$$

$$g_i^l = \sum_i w_{j,i}^{l+1} g_i^l \quad (6)$$

where x_i^l and g_i^l are the activation and gradient of i th neuron at l th layer, $w_{j,i}^{l+1}$ denotes the weight of neuron i at l -th layer and neuron j at $l+1$ -th layer. Different CNN architectures have arisen in recent research growths. AlexNet has been used in this work. For the 2012 ImageNet competition, it was implemented to lower the picture error in classification from 26 to 15.3%. It's an incredibly competent and well-organized architecture. Its eight learning layers comprise three completely connected layers and five convolution layers. To construct the class labels, the output of the last layer is input into the softmax activation function. GPU sharing connects the second, fourth, and fifth levels' kernels to their preceding layers. The second layer and the third layer kernel are entirely connected. The max-pooling layers are connected to the normalization layer after the first and second layers. Each learning layer is associated with ReLU activation function. The network architecture details are shown in Table 1. The neurons in the last layers are set to 22 to balance the features. The output layer, layer 12, has a sigmoid activation function that indicates the efficient properties of waste goods. This layer is provided as input to the DBN, which classifies waste products into recyclable and non-recyclable categories.

3.3.2 VO Sequencing Using LSTM

The backbone network in this work was a dense layer LSTM. Figure 4 displays the three layers of the thick LSTM. It consists of two fully connected (FC) layers; the first has 160 neurons, and the second has 90 neurons. The following layers are batch normalisation and dropout. The final layer is the FC, which has three neurons used to segment the picture. The dense layer comes after the LSTM to divide the area around the brain tumour. Features are fed into the LSTM layer from the ROI and G-CA, and CNN. The dataset's maximum

Table 1 CNN architecture details

Layer number	Layers	Size
0	Input preprocessed image	1280×384
1	Convolution_64 filters	640×192
2	Max pooling	320×96
3	Convolution_128 filters	160×48
4	Max pooling	80×24
5	Convolution_256 filters	40×12
6	Convolution_384 filters	40×12
7	Convolution_512 filters	40×12
8	Max pooling	20×6
9	Fully connected	512 neurons
10	Fully connected	512 neurons
11	Fully connected	512 neurons

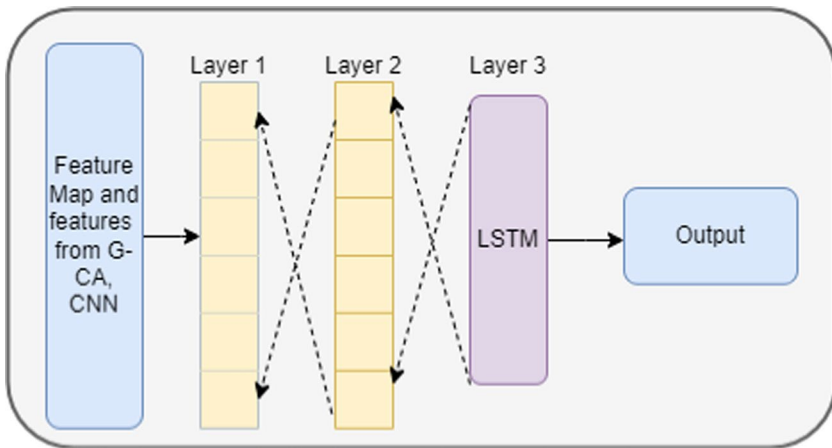


Fig. 4 Dense LSTM structure

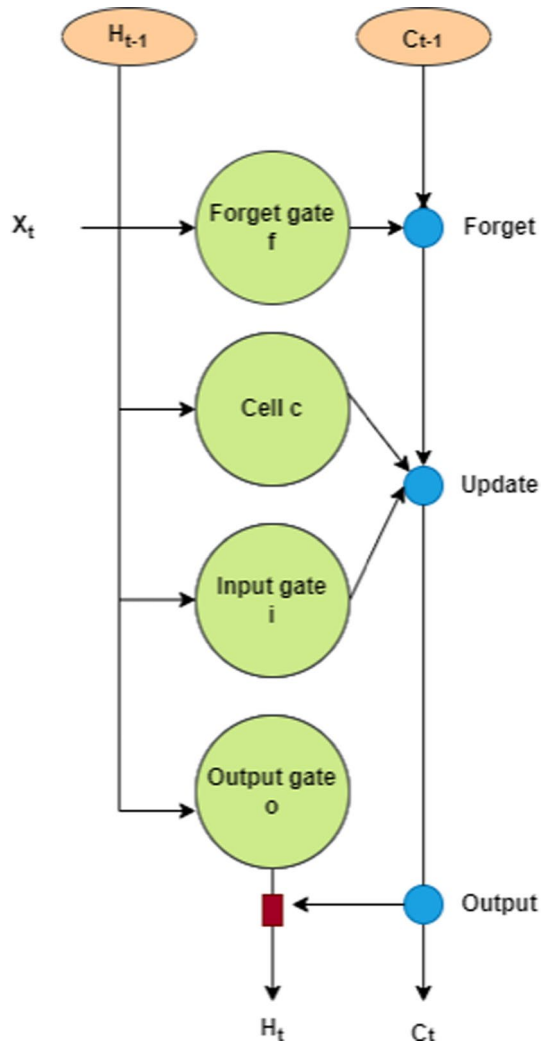
number of slices is 30, which equals the number of sequences defined. There are 225 hidden units in the first and third layers of the LSTM and 200 hidden units in the second and fourth layers. As seen in Fig. 5, each layer was made up of LSTM units with four gates, such as input (i), forget (f), cell (c), and output (o).

In Fig. 5, the variables X, C and H declares the input, cell and hidden states sequentially. In each LSTM block, three weights such as input weight I_w , recurrent weight R_w and bias b has been used as in Eq. (7)

$$I_w = \begin{bmatrix} I_{oi} \\ I_{of} \\ I_{oc} \\ I_{oo} \end{bmatrix}, R_w = \begin{bmatrix} R_{oi} \\ R_{of} \\ R_{oc} \\ R_{oo} \end{bmatrix}, \text{ and } b = \begin{bmatrix} b_i \\ b_f \\ b_c \\ b_o \end{bmatrix} \tag{7}$$

The cell state at certain time step t is declared as follows,

Fig. 5 LSTM block



$$C_t = F_t \odot C_{t-1} + i_t \odot c_t \quad (8)$$

where \odot is the Hadamard product. The concealed state H_t of t is denoted as,

$$H_t = o_t \odot \tanh(C_t) \quad (9)$$

3.4 Cost Function Optimization

Consider using the suggested C-LSTM based VO system to calculate the conditional probability of the poses $y_t = (y_1, y_2, \dots, y_t)$ with the RGB sequential of monocular images $x_t = (x_1, x_2, \dots, x_t)$ up to the time t in the probabilistic form:

$$p(y_t|x_t) = p(y_1, y_2, \dots, y_t | x_1, x_2, \dots, x_t) \quad (10)$$

The C-LSTM is used for both probabilistic inference and modeling. The DNN maximizes in order to determine the ideal parameters θ^* for the VO:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} p(y_t|x_t; \theta) \quad (11)$$

4 Results and Discussion

This section uses the popular KITTI VO/SLAM benchmark to assess the suggested end-to-end monocular VO approach [42]. Most currently available monocular video encoding techniques do not compute an absolute scale, so their localization outcomes must be manually matched to the actual data. As a result, the open-source VO library LIBVISO2 [43] is used for comparison. It recovers the scale for the monocular VO using a fixed camera height. It also uses its stereo version, which may acquire the absolute positions directly.

4.1 Dataset

There are 22 image sequences in the KITTI VO/SLAM benchmark [42], 11 of which (Sequence 00–10) are linked to ground truth. The remaining ten sequences (Sequences 11–21) merely contain raw sensor data. The fact that this dataset was captured at a relatively slow frame rate (10 frames per second) while driving through crowded, dynamic cities at speeds of up to 90 km/h makes it extremely difficult for monocular VO algorithms to process.

Two different experiments were carried out to assess the suggested approach. As ground truth is only available for these sequences, the first one is based on Sequence 00–10 to analyse its performance statistically. The relatively long sequences 00, 02, 08, and 09 are the only ones utilised for training to create a separate dataset for testing. The paths are divided into segments of varying lengths to produce large training data—7410 samples altogether. The tested, trained models are evaluated on the following sequence: 03, 04, 05, 06, 07, and 10. As the capacity to extrapolate effectively to actual data is crucial for deep learning methods, the subsequent trial examines the suggested technique's behaviour and the trained VO models in entirely novel settings. This is also necessary for the VO problem, as previously explained. As a result, models trained on all of Sequence 00–10 are tested on Sequence 11–21, which lacks training ground truth.

The network is trained using an NVIDIA Tesla K40 GPU based on the well-known DL framework Theano. It is trained using the Adagrad optimiser for a maximum of 200 epochs at a learning rate of 0.001. Techniques such as dropout and early halting are implemented to prevent the models from overfitting. The CNN is based on a pre-trained FlowNet model to minimise both the training time and data required to converge [44].

4.2 Experimental Results of VO

The KITTI VO/SLAM evaluation metrics, which calculate the average root mean square errors (RMSEs) of translational and rotating errors for all sequences of lengths between

100 and 800 m and various speeds (the range of speeds varies in different sequences), are used to analyse the performance of the trained VO models.

Sequences 00, 02, 08, and 09 are used to train the initial DL-based model. Sequences 03 to 07 and 10 are used for testing. In Fig. 6, the translation and rotation against various path lengths and speeds are displayed along with the average RMSEs of the calculated VO on the test sequences. Due to the implementation of 6G network, the high drifts are avoided and the proposed model secured improved results than stereo VISO2-S, Monocular VISO2-M and DeepVO [45]. The rotational errors are smaller than the translation errors since the KITTI dataset is recorded while the car is moving, which tends to be high speed on driving and slow in rotation with varied velocity. As seen in Fig. 6a, b while the trajectory length is increased, the translation and rotation errors are reduced compared to the cases considered, such as stereo, monocular, and DeepVO. Also, in Fig. 7a, b the translation error and rotation error decrease as speed increases.

Table 2 summarises the detailed performance of the algorithms on the testing sequences. It suggests that compared to the examined VO systems, the C-LSTM produces more reliable results. While the previous experiment assessed the generalisation of the proposed model, the network is tested on the KITTI VO benchmark testing dataset to explore its performance in entirely new settings with distinct motion patterns and images. The KITTI VO benchmark's 11 training sequences, or Sequence 00–10, are used to train the C-LSTM model. This provides additional data to minimise overfitting and optimise the network's performance. The VO findings cannot be subjected to any quantitative analysis because no ground truth is available for these testing sessions.

- T_{rel} -Average translation RMSE (%) of 100–800 m length.
- R_{rel} is the average rotational RMSE ($^{\circ}/100$ m) for 100–800 m length.

The C-LSTM VO produces results that are substantially superior to those of the monocular VISO2 and somewhat comparable to those of the stereo VISO2. It appears that this larger training dataset improves the performance of the proposed model. The DeepVO, a monocular VO method, provides an attractive performance, demonstrating that the trained model may generalise effectively in new settings, considering the stereo features of the

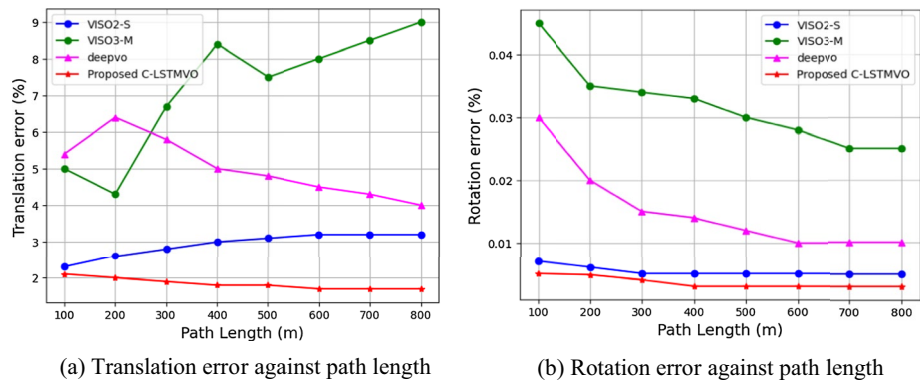


Fig. 6 Error calculation during fixed-length path travel **a** Translation error against path length **b** Rotation error against path length

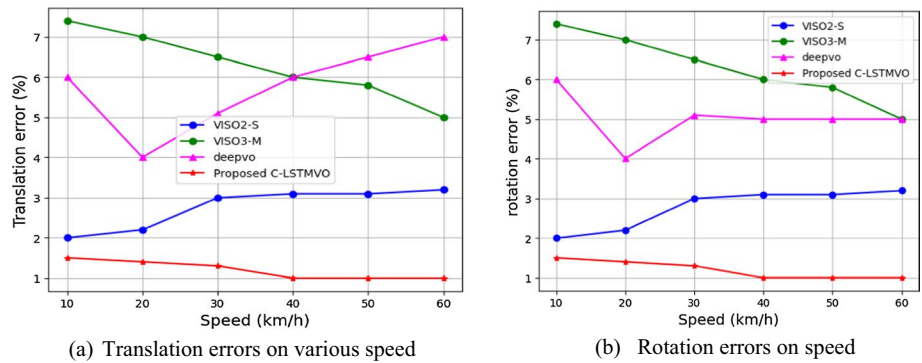


Fig. 7 Error calculation during various speeds of path travel **a** Translation errors on various speed **b** Rotation errors on speed

Table 2 Testing sequence results

Seq. No	VISO2-S		VISO2-M		DeepVO		C-LSTMVO (Proposed)	
	T_{rel} (%)	R_{rel} (°)	T_{rel} (%)	R_{rel} (°)	T_{rel} (%)	R_{rel} (°)	T_{rel} (%)	R_{rel} (°)
03	3.2	3.34	8.45	8.76	8.26	6.26	2.87	2.91
04	2.11	2.11	4.67	4.48	7.16	6.15	1.98	2.1
05	1.52	1.59	18.21	17.54	2.61	3.21	1.34	1.42
06	1.48	1.46	7.28	6.13	5.32	5.78	1.35	1.39
07	1.86	1.9	23.61	28.76	3.81	4.24	1.76	1.82
10	1.11	1.20	40.23	31.98	8.12	8.23	1.02	1.1
Mean	1.88	1.93	17.075	16.28	5.88	5.65	1.72	1.79

stereo VISO2. One possible exception is the Sequence 10 test, which has quite large localisation errors while having a trajectory shape that is similar to the stereo VISO2s. There are multiple causes. Firstly, there is insufficient data at high speeds in the training dataset. Only Sequence 01, out of the 11 training datasets, exhibits velocities greater than 60 km/h. On the other hand, Sequence 10's top speeds range from 50 to around 90 km/h. Furthermore, only 10 Hz is used to collect the pictures, which increases the difficulty of VO estimate during rapid movement.

5 Conclusion

This work presents an innovative deep learning-based end-to-end monocular video algorithm. This new paradigm combines CNNs with LSTM to achieve simultaneous representation training and sequential monocular voice-over network (VO) modelling, leveraging the power of GCA-CNN and LSTM. There is no need to properly adjust the VO system's parameters because it is trained end-to-end and does not rely on any module in the traditional VO algorithms—not even camera calibration—for posture estimation. It is confirmed

that it can generate exact VO findings with exact scales and function well in new contexts based on the KITTI VO benchmark. The Analyzed results with the comparison among the considered VO approaches show the efficiency of the proposed model with reduced error rate on both testing and training video sequences. Quite the contrary: by combining geometry with the representation, knowledge, and models that the DNNs have learned, it can be a useful supplement, helping to further enhance the VO's accuracy and, more importantly, robustness.

Acknowledgements Not applicable.

Author Contributions Cheng Zhang: Conceptualization, Methodology, Formal analysis, Validation, Resources, Supervision, Writing—original draft, Writing—review and editing. Yuchan Yang: Resources, Supervision, Writing—original draft, Writing—review and editing. Guangyao Li: Resources, Supervision, Writing—review and editing.

Funding Not applicable.

Data Availability The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethics Approval Not applicable.

Consent for Publication Not applicable.

References

- Zhang, Y., Wu, Y., Tong, K., Chen, H., & Yuan, Y. (2023). Review of visual simultaneous localization and mapping based on deep learning. *Remote Sensing*, *15*, 2740.
- Scaramuzza, D., & Fraundorfer, F. (2011). Visual odometry: Tutorial. *IEEE Robotics & Automation Magazine*, *18*(4), 80–92.
- Qu, J., Mao, B., Li, Z., Xu, Y., Zhou, K., Cao, X., & Wang, X. (2023). Recent progress in advanced tactile sensing technologies for soft grippers. *Advanced Functional Materials*. <https://doi.org/10.1002/adfm.202306249>
- Chen, J., Wang, Q., Cheng, H. H., Peng, W., & Xu, W. (2022). A review of vision-based traffic semantic understanding in itss. *IEEE Transactions on Intelligent Transportation Systems*, *23*(11), 19954–19979.
- Bala, J. A., Adeshina, S. A., & Aibinu, A. M. (2022). Advances in visual simultaneous localisation and mapping techniques for autonomous vehicles: A review. *Sensors*, *22*, 8943. <https://doi.org/10.3390/s22228943>
- Chen, J., Xu, M., Xu, W., Li, D., Peng, W., & Xu, H. (2023). A flow feedback traffic prediction based on visual quantified features. *IEEE Transactions on Intelligent Transportation Systems*, *24*(9), 10067–10075.
- Chen, J., Wang, Q., Peng, W., Xu, H., Li, X., & Xu, W. (2022). Disparity-based multiscale fusion network for transportation detection. *IEEE Transactions on Intelligent Transportation Systems*, *23*(10), 18855–18863.
- Peng, J., Hou, Y., Xu, H., & Li, T. (2022). Dynamic visual SLAM and MEC technologies for B5G: A comprehensive review. *EURASIP Journal on Wireless Communications and Networking*. <https://doi.org/10.1186/s13638-022-02181-9>
- Li, S., Chen, J., Peng, W., Shi, X., & Bu, W. (2023). A vehicle detection method based on disparity segmentation. *Multimedia Tools and Applications*, *82*(13), 19643–19655.
- Xu, B., & Guo, Y. (2022). A novel DVL calibration method based on robust invariant extended Kalman filter. *IEEE Transactions on Vehicular Technology*, *71*(9), 9422–9434.

11. Xu, B., Wang, X., Zhang, J., Guo, Y., & Razzaqi, A. A. (2022). A novel adaptive filtering for cooperative localization under compass failure and non-gaussian noise. *IEEE Transactions on Vehicular Technology*, 71(4), 3737–3749.
12. Sun, R., Dai, Y., & Cheng, Q. (2023). An adaptive weighting strategy for multisensor integrated navigation in urban areas. *IEEE Internet of Things Journal*, 10(14), 12777–12786.
13. Lei, J., Fang, H., Zhu, Y., Chen, Z., Wang, X., Xue, B., & Wang, N. (2024). GPR detection localization of underground structures based on deep learning and reverse time migration. *NDT & E International*, 143, 103043.
14. Jiang, H., Wang, M., Zhao, P., Xiao, Z., & Dustdar, S. (2021). A utility-aware general framework with quantifiable privacy preservation for destination prediction in LBSS. *IEEE/ACM Transaction on Networking*, 29(5), 2228–2241.
15. Liu, D., Cao, Z., Jiang, H., Zhou, S., Xiao, Z., & Zeng, F. (2022). Concurrent low-power listening: A new design paradigm for duty-cycling communication. *ACM Transactions on Sensor Networks*, 19(1), 1–24.
16. Jiang, H., Chen, S., Xiao, Z., Hu, J., Liu, J., & Dustdar, S. (2023). Pa-count: Passenger counting in vehicles using wi-fi signals. *IEEE Transactions on Mobile Computing*. <https://doi.org/10.1109/TMC.2023.3263229>
17. Xiao, Z., Fang, H., Jiang, H., Bai, J., Havyarimana, V., Chen, H., & Jiao, L. (2023). Understanding private car aggregation effect via spatio-temporal analysis of trajectory data. *IEEE Transactions on Cybernetics*, 53(4), 2346–2357.
18. Xiao, Z., Li, H., Jiang, H., Li, Y., Alazab, M., Zhu, Y., & Dustdar, S. (2023). Predicting urban region heat via learning arrive-stay-leave behaviors of private cars. *IEEE Transactions on Intelligent Transportation Systems*, 24(10), 10843–10856.
19. Ma, S., Chen, Y., Yang, S., Liu, S., Tang, L., Li, B., & Li, Y. (2023). The autonomous pipeline navigation of a cockroach bio-robot with enhanced walking stimuli. *Cyborg and Bionic Systems*, 4, 0067.
20. Fu, C., Yuan, H., Xu, H., Zhang, H., & Shen, L. (2023). TMSO-Net: Texture adaptive multi-scale observation for light field image depth estimation. *Journal of Visual Communication and Image Representation*, 90, 103731.
21. Liu, H., Yuan, H., Hou, J., Hamzaoui, R., & Gao, W. (2022). PUFA-GAN: A frequency-aware generative adversarial network for 3d point cloud upsampling. *IEEE Transactions on Image Processing*, 31, 7389–7402.
22. Wu, Z., Zhu, H., He, L., Zhao, Q., Shi, J., & Wu, W. (2023). Real-time stereo matching with high accuracy via Spatial Attention-Guided Upsampling. *Applied Intelligence*, 53(20), 24253–24274.
23. Zhang, C., Zhou, L., & Li, Y. (2024). Pareto optimal reconfiguration planning and distributed parallel motion control of mobile modular robots. *IEEE Transactions on Industrial Electronics*, 71(8), 9255–9264.
24. Zhao, Y., Chen, S., Liu, S., Hu, Z., & Xia, J. (2024). Hierarchical equalization loss for long-tailed instance segmentation. *IEEE Transactions on Multimedia*. <https://doi.org/10.1109/TMM.2024.3358080>
25. Gu, Y., Hu, Z., Zhao, Y., Liao, J., & Zhang, W. (2024). MFGTN: A multi-modal fast gated transformer for identifying single trawl marine fishing vessel. *Ocean Engineering*, 303, 117711.
26. Chen, Y., Li, N., Zhu, D., Zhou, C. C., Hu, Z., Bai, Y., & Yan, J. (2024). BEVSOC: Self-supervised contrastive learning for calibration-free bev 3d object detection. *IEEE Internet of Things Journal*. <https://doi.org/10.1109/JIOT.2024.3379471>
27. Qi, F., Tan, X., Zhang, Z., Chen, M., Xie, Y., & Ma, L. (2024). Glass makes blurs: Learning the visual blurriness for glass surface detection. *IEEE Transactions on Industrial Informatics*, 20(4), 6631–6641.
28. Zou, W., Sun, Y., Zhou, Y., Lu, Q., Nie, Y., Sun, T., & Peng, L. (2022). limited sensing and deep data mining: A new exploration of developing city-wide parking guidance systems. *IEEE Intelligent Transportation Systems Magazine*, 14(1), 198–215.
29. Cheng, B., Wang, M., Zhao, S., Zhai, Z., Zhu, D., & Chen, J. (2017). Situation-aware dynamic service coordination in an IoT environment. *IEEE/ACM Transactions on Networking*, 25(4), 2082–2095.
30. Zheng, W., Lu, S., Yang, Y., Yin, Z., Yin, L., & Ali, H. (2024). Lightweight transformer image feature extraction network. *PeerJ Computer Science*, 10, e1755.
31. Mi, C., Liu, Y., Zhang, Y., Wang, J., Feng, Y., & Zhang, Z. (2023). A vision-based displacement measurement system for foundation pit. *IEEE Transactions on Instrumentation and Measurement*, 72, 2525715.
32. Zhang, H., Luo, G., Li, J., & Wang, F. Y. (2022). C2FDA: Coarse-to-fine domain adaptation for traffic object detection. *IEEE Transactions on Intelligent Transportation Systems*, 23(8), 12633–12647.

33. Sheng, H., Wang, S., Yang, D., Cong, R., Cui, Z., & Chen, R. (2023). Cross-view recurrence-based self-supervised super-resolution of light field. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12), 7252–7266.
34. Tustison, N. J., et al. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29, 1310–1320.
35. Yin, F., Lin, Z., Kong, Q., Xu, Y., Li, D., Theodoridis, S., & Cui, S. R. (2020). FedLoc: Federated learning framework for data-driven cooperative localization and location data processing. *IEEE Open Journal of Signal Processing*, 1, 187–215.
36. Xu, G., Zhang, Q., Song, Z., & Ai, B. (2023). Relay-assisted deep space optical communication system over coronal fading channels. *IEEE Transactions on Aerospace and Electronic Systems*, 59(6), 8297–8312.
37. Simonyanand, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXivpreprint arXiv:1409.1556*
38. Tan, J., Zhang, K., Li, B., & Wu, A. (2023). Event-triggered sliding mode control for spacecraft reorientation with multiple attitude constraints. *IEEE Transactions on Aerospace and Electronic Systems*, 59(5), 6031–6043.
39. Di, Y., Li, R., Tian, H., Guo, J., Shi, B., Wang, Z., Liu, Y., (2023). A maneuvering target tracking based on fastIMM-extended Viterbi algorithm. *Neural Computing and Applications*.
40. Zhao, S., Liang, W., Wang, K., Ren, L., Qian, Z., Chen, G., & Ren, L. (2024). A multiaxial bionic ankle based on series elastic actuation with a parallel spring. *IEEE Transactions on Industrial Electronics*, 71(7), 7498–7510.
41. Wang, K., Williams, H., Qian, Z., Wei, G., Xiu, H., Chen, W., & Ren, L. (2023). Design and evaluation of a smooth-locking-based customizable prosthetic knee joint. *Journal of Mechanisms and Robotics*, 16(4), 041008.
42. Cai, L., Yan, S., Ouyang, C., Zhang, T., Zhu, J., Chen, L., & Liu, H. (2023). Muscle synergies in joystick manipulation. *Frontiers in Physiology*, 14, 1282295.
43. Khan, D., Alonazi, M., Abdelhaq, M., Al Mudawi, N., Algarni, A., Jalal, A., & Liu, H. (2024). Robust human locomotion and localization activity recognition over multisensory. *Frontiers in Physiology*, 15, 1344887.
44. Wang, F., Ma, M., & Zhang, X. (2024). Study on a portable electrode used to detect the fatigue of tower crane drivers in real construction environment. *IEEE Transactions on Instrumentation and Measurement*, 73, 2506914.
45. He, H., Chen, Z., Liu, H., Liu, X., Guo, Y., & Li, J. (2023). Practical tracking method based on best buddies similarity. *Cyborg and Bionic Systems*, 4, 50.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Cheng Zhang Senior Engineer at Lotus Technology Co., Ltd. Master's in Business Administration from UESTC, Bachelor's in Digital Media from Nanjing University. Extensive experience includes senior roles at Geely Automotive Research Institute and Tianzong Network Technology Co., Ltd. Expertise in intelligent driving technology. Notable contributions include research publications and patents in autonomous driving.



Yuchan Yang Engineer at Lotus Technology Co., Ltd. Master's in Software Engineering from Shandong Academy of Sciences and Shanxi University. With expertise in deep learning frameworks, she has led projects in 5G shared mobility and domain adaptation for image recognition. Her achievements include research on 3D map construction for autonomous driving.



Guangyao Li is a male associate professor at Tongji University, specializing in Electronic and Information Engineering. With a Ph.D. in Aerospace Manufacturing and extensive experience in CAD research. He has led several research projects funded by the National Natural Science Foundation of China, focusing on image restoration and MRI algorithms. Li has authored numerous publications in prestigious journals and conferences, including research on super-resolution imaging and scene text detection. He is recognized for his expertise in image processing, computer vision, and artificial intelligence.