IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 February 2023, accepted 26 February 2023, date of publication 2 March 2023, date of current version 8 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3251733

RESEARCH ARTICLE

A Robust Visual SLAM Method for Additive Manufacturing of Vehicular Parts Under Dynamic Scenes

WENBO XU[1,2], WEIWEI FAN[1,2], JINGYANG LI[3], OSAMA ALFARRAJ[4], AMR TOLBA[4], (Senior Member, IEEE), AND TIANHONG HUANG[5]

[1]School of Vehicle and Traffic Engineering, Henan Institute of Technology, Xinxiang 453003, China
[2]Henan Engineering Research Center of NVH Control for New-Energy Vehicle, Xinxiang 453003, China
[3]School of Mechanical Engineering, Henan Institute of Technology, Xinxiang 453003, China
[4]Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia
[5]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, USA

Corresponding author: Wenbo Xu (xwb@hait.edu.cn)

This work was funded by the Researchers Supporting Project Number (RSP2023R102), King Saud University, Riyadh, Saudi Arabia. It was also supported in part by the Ministry of Education's First Batch of Industry University Cooperation Collaborative Education Projects, in 2022, under Grant 220602436070556; in part by the Henan Research Program of Science and Technology under Grant 212102311151; in part by the Ministry of Education's Second Batch of Industry University Cooperation Collaborative Education Projects, in 2021, under Grant 202102327002, in part by the Ministry of Education's First Batch of Industry University Cooperation Collaborative Education Projects, in 2021, under Grant 202101014006.

ABSTRACT Additive manufacturing has significant advantages in complex parts of the vehicle manufacturing. As additive manufacturing is a kind of precise production activity, different components of manufacturing instruments need to be located in appropriate positions to ensure accuracy. The visual Simultaneous Localization and Mapping (SLAM) can be considered to be a practical means for this purpose. Considering dynamic characteristics of additive manufacturing scenarios, this paper constructs a deep learning-enhanced robust SLAM approach for production monitoring of additive manufacturing. The proposed method combines the semantic segmentation technique with the motion-consistency detection algorithm together. Firstly, the Transformer-based backbone network is used to segment the images to establish the a prior semantic information of dynamic objects. Next, the feature points of dynamic objects are projected by the motion-consistency detection algorithm. Then, the static feature points are adopted for feature matching and position estimation. In addition, we conducted a couple of experiments to test function of the proposed method. The obtained results show that the proposal can have excellent performance to promote realistic additive manufacturing process. As for numerical results, the proposal can improve image segmentation effect about 10% to 15% in terms of scenarios of visual SLAM-based additive manufacturing.

INDEX TERMS Additive manufacturing, vehicular parts, visual SLAM, deep learning, dynamic scenes.

I. INTRODUCTION

Additive Manufacturing is an emerging processing technology based on the principle of discrete stacking [1], [2], [3], which breaks the traditional reduced material manufacturing and equal material manufacturing production methods [4], [5]. It is a new manufacturing technology that does not require the collaboration of jigs and fixtures and is not processed by machine tools and equipment [6], [7], [8]. With the unification of production standards and the maturity of raw material technology, additive manufacturing technology is well boosted by intelligent technology and the intersection of basic disciplines like automotive industry, aerospace, bioengineering and other fields [9], [10]. At the same time, its gradual prevalence can also breed some latent technical breakthroughs in many cross-discipline applications [11], [12]. Therefore, it is believed to have unlimited market potential in terms of smart manufacturing [13], [14], [15].

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Quan.

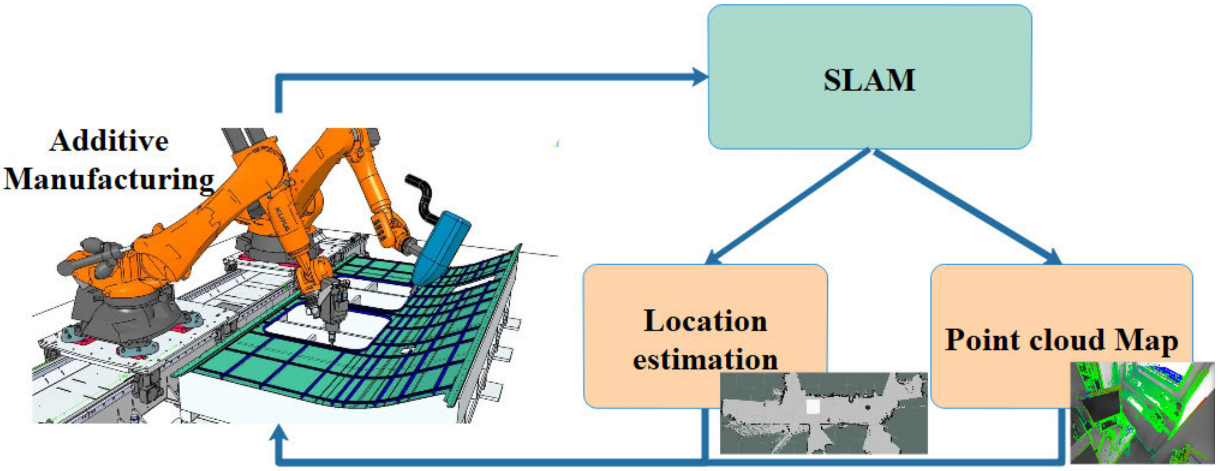


FIGURE 1. A typical example that illustrates scenarios of robust visual SLAM for additive manufacturing.

Additive manufacturing technology firstly uses computer-aid design softwares to conduct 3D modeling for mechanical parts. Then, slicing software is utilized to slice the 3D models according to the parameters of the parts. On this basis, the computer algorithms are used to precisely connect each layer to form a layer stack to quickly realize the additive manufacturing of parts [16], [17]. Contemporarily, the application of additive manufacturing technology has bee more and more general in the automotive industry [18], [19],liu2022human. In this context, some brand-name automotive companies currently choose to use additive manufacturing technology in the automotive development stage to achieve the purpose of rapid verification and optimization of components [20], [21]. In the small batch production of automotive parts often involve some complex parts with thin walls and internal abdominal cavities, the traditional forging and casting processes have limitations in processing and cannot meet the production requirements [22], [23]. Due to the point-by-point, line-by-line and domain-by-domain local forming characteristics of additive manufacturing technology, it is possible to achieve highly flexible near-net-shape additive manufacturing in the manufacture of complex parts [24], [25]. Therefore, additive manufacturing technology has significant advantages in the manufacturing of complex parts, and the application prospect is very promising [26], [27].

A. MOTIVATION

In additive manufacturing, how to use vision sensors for accurate positioning and mapping of parts is the key to realize autonomation. With the continuous development of research, robots are equipped with more diverse sensors, including vision, laser, radar, and multi-sensor fusion methods. Robots are able to perceive their environment and have the ability to estimate the state of their systems using the sensors they carry, they can sense their surroundings and make decisions autonomously. These digital technologies require accurate and robust localization with the ability to progressively build and maintain models of the world scenes. In this work, localization refers to the ability to obtain the internal system state of the robot's motion, including position, orientation, and velocity. While mapping refers to the ability to sense the state of the external environment and capture information about the surroundings, including the geometry, appearance, and semantic information of a 2D or 3D scene [28]. These components can perceive internal or external states individually or, like simultaneous localization and mapping (SLAM) [29], so as to facilitate control decision of robot' poses.

The localization and mapping problem has been studied for decades and various sophisticated hand-designed (hand-designed) models and algorithms are being developed, such as odometer estimation, image-based localization, position recognition, SLAM, motion reconstruction (SfM) [29], [30]. Under ideal conditions, these sensors and models are able to estimate the system state accurately regardless of the time, environment constraints. However, in reality, sensor measurement errors, system modeling errors, complex environmental dynamics and unrealistic constraints (conditions) affect the accuracy and reliability of manually designed systems [31]. Although modern vision SLAM systems are quite mature and have satisfactory performance [32], the aforementioned classical SLAM systems are with the assumption that the objects for SLAM are static, and the detection and processing of dynamic objects are very limited.

However, in actual indoor and outdoor scenes, it is impossible to circumvent moving objects [33]. In this case, unexpected changes in the surrounding environment may seriously affect the camera pose estimation, increase the trajectory error or even lead to system failure. Thus, the detection of moving objects and the correct segmentation of dynamic regions become important research aspects of vision SLAM in dynamic scenes. Because of the limitations of model-based solutions and the rapid development of machine learning, especially deep learning, researchers have been prompted to consider data-driven learning methods as an alternative approach to solve this issue. The class of relationships between sensor data input values (e.g., vision, inertial guidance, LiDAR data or other sensors) and target output values (e.g., position, orientation, scene geometry or semantics) as a mapping function [34], [35].

B. CONTRIBUTIONS

While traditional model-based solutions are implemented by manually designing algorithms, learning-based approaches construct this mapping function by learning large amounts of data. The learning-based approach has three advantages. Firstly, the learning approach can automatically discover task-relevant features using a highly expressive deep neural network as a general-purpose approximator. This feature enables trained models to adapt to various scenarios (e.g., featureless scenes, dynamic high-speed scenes, dynamic blur, accurate camera calibration) [36], [37]. Secondly, the learning approach allows learning from past experiences and actively developing new information. By building a general data-driven model, researchers can solve domain-specific problems without having to go through the trouble of specifying the entire knowledge about mathematical and physical rules when building the model. Thirdly, deep neural networks have the ability to be scaled to large-scale problems. Trained on large data sets through back propagation and gradient descent algorithms, a large number of parameters in the DNN framework can be automatically optimized by minimizing the loss function. Thus, harnessing the power of data and computation to solve localization and mapping is potentially achievable.

The multi-sensor fusion scheme needs to add information from different sensor sources, facing multiple difficulties such as data correlation, signal synchronization, and fusion processing, which greatly increases the complexity of the system, and the dense scene flow approach is computationally intensive and a great challenge for real-time computing [38], [39]. The multi-sensor fusion scheme can construct semantic maps to enrich the robot's understanding of the environment and thus obtain advanced perception, but there are problems of misjudgment for movable objects. To address it, this paper proposes a robust SLAM algorithm for dynamic scenes, which uses deep learning to quickly identify dynamic object frames combined with sparse feature optical flow calculation to make further dynamic judgments, the scenario of the proposed method is shown in Figure 1. Edge detection algorithms are used to effectively segment the edges of dynamic objects to ensure that no too many static feature points are mistakenly removed. And a static environment 3D point cloud map without dynamic objects are constructed to truly realize the powerful sensing capability of autonomous robots.

To sum up, the main contributions of this paper can be stated as the following three aspects:

- This work aims at the additive manufacturing of automobile parts, and explores to employ deep learning-based vision sensing to enhance the manufacturing process.
- This work proposes a robust visual SLAM method for additive manufacturing of vehicular parts under dynamic scenes.
- This work conducts simulation experiments on real-world scenarios to evaluate performance of the proposal, and corresponding discussions are also made for it.

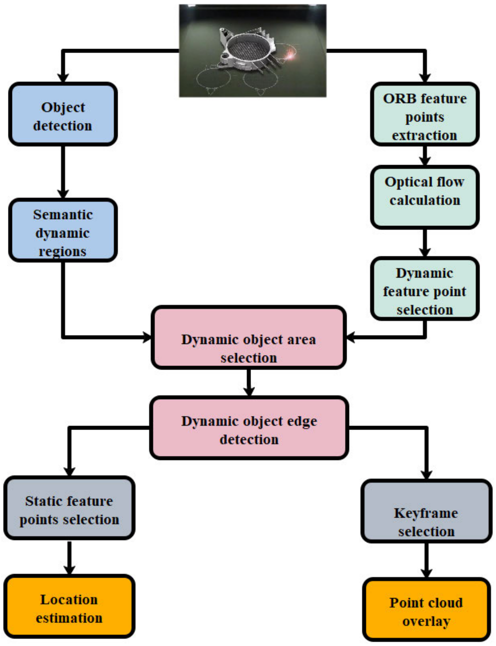

Object detection
ORB feature points extraction
Semantic dynamic regions
Optical flow calculation
Dynamic feature point selection
Dynamic object area selection
Dynamic object edge detection
Static feature points selection
Keyframe selection
Location estimation
Point cloud overlay

FIGURE 2. The workflow for main detailed process of proposed methodology.

II. METHODOLOGY

In our study, the transformer real-time target detection algorithm is used to quickly obtain the rough rectangular range of potential semantic dynamic objects in the three-channel image of input, the ORB feature points and the optical flow field are extracted and calculated, respectively, which largely reduce the time to calculate the optical flow field of all pixel points. Then by combining the semantic data with the dynamic feature points filtered by the optical flow field calculation, the true motion of the object can be obtained. Then, the canny operator is adopted to detect the edges of the dynamic objects to extract the edge data of the dynamic objects, and to do position estimation of camera by minimizing re-projection error of static feature points other than dynamic objects. Finally, the map is constructed using the key frames with the dynamic objects removed. The overall flow is shown in Figure 2.

A. REAL-TIME TARGET DETECTION BASED ON TRANSFORMER

The Detection Transformer (DEtection TRansformer, DETR) [40], [41] with an ensemble global loss that makes predictions through bilateral match and a classical encoder-decoder architecture, which containing three components:

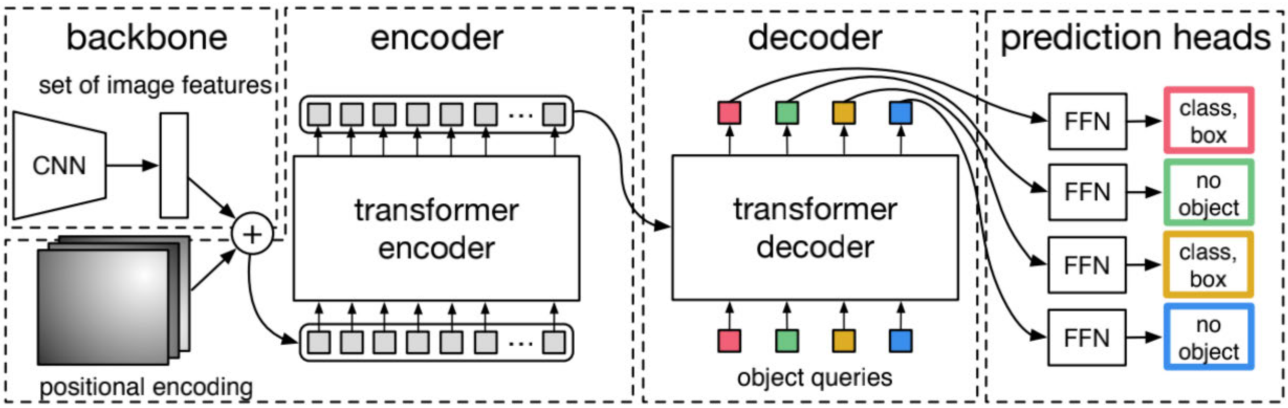


FIGURE 3. Sketch map for the technical structure of a Transformer-based vision sensing approach.

a CNN based backbone to extracte feature representations, a Transformer pretraining model to enhance features, and a simple feedforward network (FFN) for performing the object detection prediction.The detail structure is shown as Figure 3. Starting from an initial image $x_{img} \in R^{3\times H_0\times W_0}$ (3 color channels, To batch the input images together with sufficient 0 padding to have the same dimension (H_0,W_0) as the largest image in same batch), a convolutional network then to generate a activation map $f \in R^{C\times H\times W}$ with lower resolution.

First, the high-level activation map of the channel dimension f is reduced from C to dimension d using a 1×1 convolution to generate a new feature map, which is written as $z_0 \in R^{d\times H\times W}$. Since a sequence is expected for the encoder as input, so the spatial dimension of the feature map z_0 is collapsed to generate a new feature map with dimension $d \times HW$. For an encoder, it is constituted by a head part, an attention mechanism part and a FFN part. Since the architecture of transformer is alignment-independent (order-independent), a fixed position encoding [8], [9] is provided and the processed results are added to the input of each attention layer.

For a decoder, it transforms N embeddings with size d by multi-head attention mechanism. The authors in [40] adopted an auto-regressive model to predict one element of the output sequence at once. Because the decoder is also permutation-independent (order-independent), hence thedifferent results will be produced according to N input embeddings. And the input embeddings are learned through the positional encodings, and the results is defined as the queries of object, and then are added to the input of each attention layer. The decoder can process N queries of objects into output feature map. The elements of the output embedding corresponding to queries of object are decoded independently into bounding box coordinates and are assigned the category labels through a FFN, which get the N final predictions. A self attention mechanism is applied on these embeddings, the model utilizes pairwise relationships between all objects to perform global inference on all objects.

The final prediction results is calculated by a three-layer foward propagation network. Also, there is a hidden layer with dimension d and a projection layer before the final results. The normalized center coordinates, height and width of the bounding box with respect to the input image are predicted by the FFN, and the category labels is predicted through a softmax function in the last layer. Thus, a fixed size of N bounding boxes is predicted, and N is typically bigger than the number of targets of interest in the original input. In addition, a category label is appended to indicate that no targets are detected within the slots (e.g., no targets of interest in the image or targets of interest do not fill the N slots). This category is similar to the ''background'' category in standard target detection methods.

B. ORB FEATURE EXTRACTION

In order to carry out the static and dynamic analysis of the object while saving the computational cost and ensuring the real time performance, this study calculates the optical flow field to estimate the motion state of the extracted ORB feature points, which are mainly divided into two parts, FAST corner point extraction and BRIEF descriptor calculation [33], which is given in the followin:

1) Construct the image pyramid, at the same time extract the FAST corner points for each pyamid layer using a uniform extraction strategy based on quadtree [41], the specific calculation process is described as follows:

Step 1: Select pixel p in the image and obtain its luminance, assumed to be Ip;

Step 2: set the threshold $T = I_p \times 0.2$;

Step 3: traverse a circle with radius 3 centered on pixel p. The 16 pixel points on the circle with radius 3;

Step 4: Let the brightness of each traversal point be I_cp. If there are N consecutive points with $I_{cp} > I_p + T$ or $I_{cp} < I_p - T$, the point is considered to be a featured point, and N is 12 in this study;

Step 5: performs the above operation for each pixel in the image.

2) Calculate the rotation angle of the FAST corner point by using the gray scale center of mass method. Define the moments of the image as:

$$m_{ab} = \sum p^a q^b \cdot I(p, q) \tag{1}$$

where $I(p, q)$ is the gray value of the FAST corner point (p, q), a, b are the order of the moments, and the image center of mass coordinates are:

$$C = (m_{10}/m_{00}, m_{01}/m_{00}) \tag{2}$$

The rotation angle is:

$$\theta = arctan(m_{10}, m_{00}) \tag{3}$$

3) Calculate the rotated BRIEF descriptor, choose the window W of $S \times S$, and define:

$$\tau\,(I; p, q) = \begin{cases} 1, & if\, I(p) < I(q) \\ 0, & else \end{cases} \tag{4}$$

where:$I(p)$ is the grayscale value at p. Randomly selected n pairs of feature points, the Generate an n-dimensional BRIEF description sub-vector:

$$f_n(w) = \sum_{1 \leq i \leq 2} 2^i \tau(I; p, q) \tag{5}$$

C. EDGE DETECTION

Directly removing the rectangular area of dynamic objects removes too much static scene, which is not conducive to accurate camera positioning and map construction. In order to extract edge of dynamic objects more accurately, this study uses canny operator to detect the edges of the filtered dynamic objects. Canny is a second-order differential operator [42], which extracts the edges of the image by the zero point of the second-order derivative at the edge of the given image. The strong edges and weak edges are detected separately, and the real weak edges can be detected. The detail steps are described as follows:

1) Eliminate image noise. Firstly, the image is smoothed by using Gaussian function. Define $f(p, q)$ as the input image, $O(x, y)$ as the output image, and $g(p, q)$ as the Gaussian function, where the Gaussian function is defined as:

$$g(p, q) = \frac{1}{2\pi\sigma^2} \exp(-\frac{p^2 + q^2}{2\sigma^2}) \tag{6}$$

$$O(p, q) = f(p, q) \times g(p, q) \tag{7}$$

2) The gradient magnitude and direction calculation. Using the image processed by Gaussian filtering, a suitable gradient operator is adopted for the gradient magnitude and direction calculation of each pixel by calculating the difference of the first-order bias between adjacent pixels. Where, A_p,A_q are the Sobel gradient operator, E_p,E_q is the difference between horizontal and vertical direction, respectively. The gradient $E(p, q)$ and direction $\theta(p, q)$ are written as following:

$$E_p(p, q) = A_p \times O(p, q) \tag{8}$$

$$E_q(p, q) = A_q \times O(p, q) \tag{9}$$

$$E(p, q) = (E_p^2 + E_q^2)^{1/2} \tag{10}$$

$$\theta(p, q) = \arctan \frac{E_q(p, q)}{E_p(p, q)} \tag{11}$$

3) Filtering non-extreme values. In the Gaussian filtering process, the edges may be amplified, and the Non-Maximum Suppression (NMS) is adopted to filter the points those are not edges. If the current calculated gradient amplitude in the field of the point is greater than along the gradient direction of the point. If the current calculated gradient amplitude of the other 2 neighboring points is greater than the gradient amplitude along the direction of the point, the point belongs to the possible edge point, otherwise it is not, and the suppression means is taken to set the gray value to 0.

4) Double threshold detection and connected edges. After the above steps of processing only get the candidate edge points, and then use the upper and lower threshold detection process to eliminate the pseudo-edge points. Points larger than. Points with upper threshold are detected as edge points, points smaller than lower threshold are detected as non-edge points, points between the two values are detected as weak edge points, and if they are adjacent to the pixel point identified as an edge point, they are judged as edge points; otherwise, they are non-edge points.

D. LOCATION ESTIMATION AND POINT CLOUD OVERLAY

After determining the exact contour of the dynamic objects, the dynamic points distributed within the objects are excluded, and only the stable feature points in the non-dynamic region are used for a more accurate camera pose solution. (u_c^i, v_c^i) is set to be the pixel coordinates of the static points in the current frame c, and the depth value z_c^i is used to obtain the 3D spatial point coordinates $P_c^i(p_c^i, q_c^i, z_c^i)$.

$$P_c^i(p_c^i, q_c^i, z_c^i) = (z_c^i \frac{u_c^i - c_p}{f_p}, z_c^i \frac{v_c^i - c_q}{f_q}, z_c^i) \tag{12}$$

where (f_p,f_q), is the focal length of camera, $(c_p, c_q,)$ is the principal point coordinates of camera.

Building 3D point cloud maps of the environment can provide better visualization of the environment. The semantic information carried by the point cloud can provide the basis for robot navigation and obstacle avoidance [22], [23]. When constructing point clouds, if there are large errors in the poses, the maps will be overlapped with obvious interlocks, which is not good for navigation. This problem can be effectively solved by overlaying the point clouds with the dynamic objects removed. The ORB_SLAM2 algorithm is used to obtain key-frames, and the point clouds of all key-frames are superimposed, which is too complicated and redundant [35], [43], [44]. In the process of key-frame screening, the following two strategies are considered: 1) key-frame validity judgment. If the area of the rejected point cloud is more than half of the current key-frame area, the key-frame is considered to contain insufficient valid information and is not involved in the overlay. 2) key-frame redundancy judgment.

TABLE 1. Display of segment performance results obtained by different methods.

Models	Pixel acc	Mean acc	Mean IoU	Frequency weight IoU
FCN-AlexNet	79.8	61.5	48.9	72.5
FCN-VGG16	77.1	57.2	52.8	63.2
Transformer(DERT)	90.2	78.3	63.5	82.2

TABLE 2. Display of running efficiency results obtained by different methods.

Systems	Number of FPs	Average time of seg	Average time of FPs extraction	Processing time per frame
ORB-SLAM2-FCN	1963	0	0.019	0.05
ORB-SLAM2-Transformer	1778	0.12	0.028	0.25

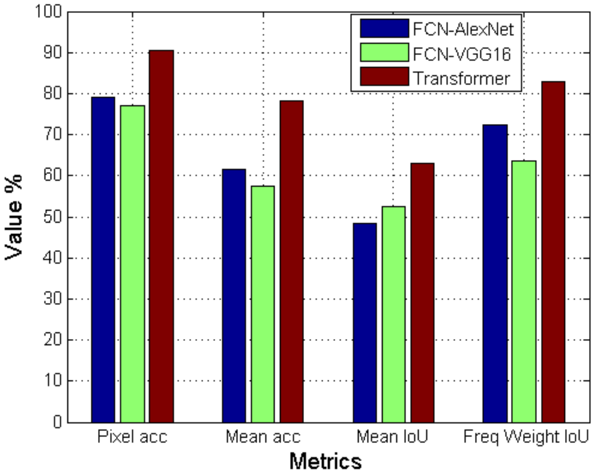


FIGURE 4. Display of segment performance results obtained by different methods.

The feature points that can be observed by multiple key-frames are called co-visual landmark points of multiple key-frames. To detect the co-viewing landmark points observed in the current key-frame, assume that the set of identified drawing key-frames is F, the set of observed landmark points is L, and the set of landmark points observed in the current key-frame is L_c, If the number of $L \cap L_c$ exceeds half of L_c, the current key-frame is considered to contain too many co-viewing landmarks and the information is redundant, so it does not participate in the superposition. If the above two conditions are satisfied, F and L_c are updated, which ensures that new point cloud information is introduced and there is enough static environment information.

III. EXPERIMENTS AND ANALYSIS

In order to evaluate the actual performance and effectiveness of the proposed ORB_SLAM2_transformer system in this paper, the system is tested in three aspects: the performance segmentation performance of the transformer network, the performance of dynamic feature point rejection, and the performance of localization in dynamic scenes.

A. EXPERIMENTAL DATA AND SETTINGS

The ‘‘Freiburg2_desk_with _person’’ dataset from the Vision Group of the Technical University of Munich (TUM), Germany, was selected as the open-source dataset, which contains a total of 4067 frames with static tables and chairs and multiple slow moving human targets [45]. This dataset is designed to check the robustness of the SLAM system to dynamic objects and people, to distinguish the map and to check the changes in the scene, which meets the requirements of the experiments in this paper.In order to test the robustness of the proposed method, an additional ‘‘DataSet_Factory’’ data set based on real scenes is constructed [36]. The data set is obtained by fixing a camera on a mobile experimental platform equipped with LIDAR, which has a static industrial assembly line and several slow-moving human targets in a total of 1715 frames, in exactly the same format as the TUM data set.

For the analysis of semantic segmentation results, this paper selects the mainstream statistical pixel accuracy (Pixel accuracy), class Mean accuracy, Mean IoU and Frequency weight IoU are the four mainstream semantic segmentation evaluation criteria used to evaluate pixel accuracy and region overlap [31]. The specific definitions are as follows:

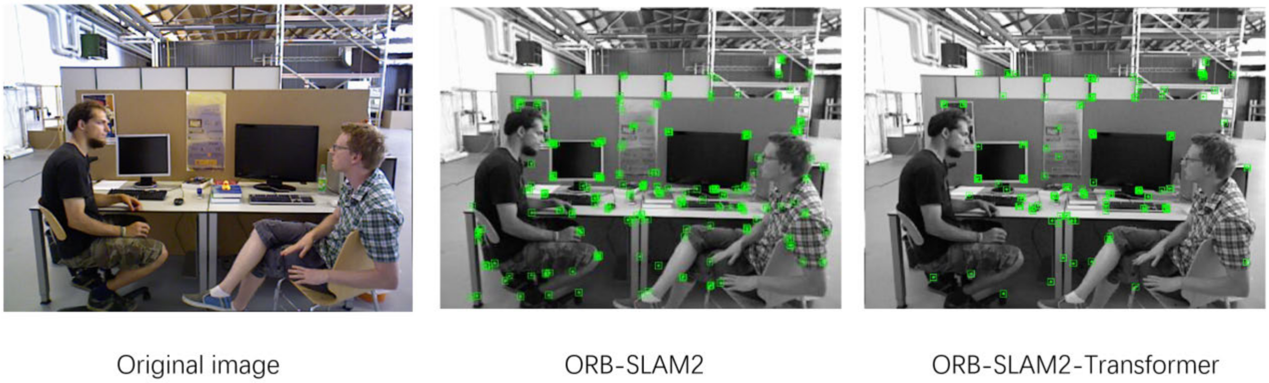
$$Pixelacc = \frac{\sum n_{ii}}{\sum t_i} \tag{13}$$

$$Meanacc = \frac{\sum \frac{n_{ii}}{t_i}}{n_{cl}} \tag{14}$$

$$MeanIoU = \sum \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \Big/ n_{cl} \tag{15}$$

$$FreqweightIoU = \sum_i \frac{t_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \Big/ \sum_k t_k \tag{16}$$

In this paper, we calculate the relative pose error (RPE) and Absolute Pose Eror (APE) to evaluate the difference

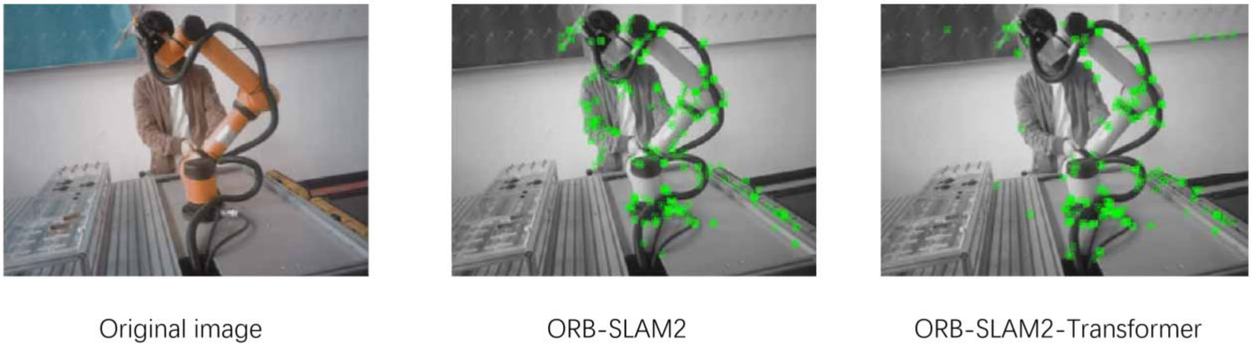


Original image

ORB-SLAM2

ORB-SLAM2-Transformer

FIGURE 5. Typical examples for the results of dynamic points projection on TUM data set.



Original image

ORB-SLAM2

ORB-SLAM2-Transformer

FIGURE 6. Typical examples for the results of dynamic points projection on DataSet_Factory data set.

in SLAM performance between the ORB_SLAM2 and the ORB_SLAM2-transformer in a dynamic environment. The relative pose error is calculated based the difference between the estimated SLAM pose and the truth value of the camera pose at the same time, and mainly describes the accuracy of the pose difference between two key frames between a fixed time Δt. The RPE for the i-th key frame is defined as:

$$E_{i:} = (Q_i^{-1} Q_{i+\Delta t})^{-1} (P_i^{-1} P_{i+\Delta t}) \tag{17}$$

where Q_i is the real trajectory pose; P_i is the key frame pose estimated by the system. Root Mean Squared Error (RMSE) is used to evaluate the error and is defined as follows:

$$RMSE(E_{i:n}, \Delta t) = (\frac{1}{m}\sum_{i=1}^{m} \|trans(E_i)\|^2)^{1/2} \tag{18}$$

In addition, one typical image segmentation method is introduced as the baseline. As this work manages to explore image segmentation-based method to enhance SLAM manufacturing process. The fully convolutional networks [46], named as FCN for short, is a most typical image segmentation method in this area. Thus, the proposal in this paper is compared with the FCN-based backbone network to measure performance appearance.

B. NUMERICAL RESULTS AND ANALYSIS

First, the transformer network based segmentation model used in this paper was analyzed. The detail performance is shown as in Table 1, and its pixel accuracy, category average accuracy and average region overlap reached 71.101%, 89.512% and 58. 157%, respectively. The performance comparison of Figure 4 indicating that this model is significantly better than other network models, and the feature map can retain more detailed features.

This is evaluated because the introduction of the transformer network semantic segmentation model increases the complexity of the system, and the consequent problem is that the feature point extraction takes longer computation time, which affects the real-time performance. In Table 2, the average feature point extraction time per image reaches 0.21757 seconds due to the addition of semantic segmentation and dynamic feature point projection algorithms to the system. Although the extraction time is significantly increased compared to the ORB_SLAM2 system, the system is still able to achieve an average rate of about 5 frames per second, which basically ensures the real-time performance.

Figure 5 shows the comparison of ORB feature point extraction results of the two systems under the TUM dataset: The feature point area contains the dynamic portrait target area; the points in the dynamic portrait target area are

TABLE 3. Comparison among experimental methods with respect to RPE performance results.

Systems	RMSE	MAX	MAE	STD
ORB-SLAM2-FCN	0.067	0.15	0.039	0.03
ORB-SLAM2-Transformer	0.050	0.12	0.038	0.025

TABLE 4. Comparison among experimental methods with respect to APE performance results.

Systems	RMSE	MAX	MAE	STD
ORB-SLAM2-FCN	0.063	0.19	0.047	0.03
ORB-SLAM2-Transformer	0.054	0.12	0.045	0.024

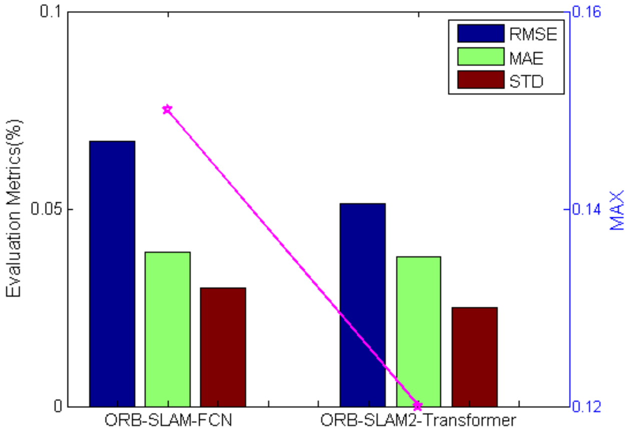


FIGURE 7. Main results about the RPE performances of experimental methods on TUM data set.

completely eliminated, and the number of eliminated feature points increases gradually as all the portrait targets enter the picture, which achieves the expected goal. Figure 6 shows the comparison of the ORB feature point extraction results of the two systems in the real scene, the feature points in the dynamic target region are completely eliminated, and the same target is achieved as the dataset, which shows that the method is still applicable in the real scene. Table 3 shows the RMSE, maximum error(MAX), mean absolute error(MAE), and standard deviation of the relative and absolute positional errors, respectively. Absolute error, and standard deviation of the relative and absolute positional errors are presented in Table 4:

As shown in Figure 7, compared with the ORB_SLAM2, the proposed ORB_ SLAM2_Transformer has a higher maximum error than the ORB_SLAM2, but the RMSE, the MAE and the standard deviation are reduced by 11.038%, 15.257% and 2. 309%, respectively; From the Figure 8, for the absolute trajectory error, the ORB_SLAM2_Transformer has a higher maximum error than the ORB_SLAM2. The ORB_SLAM2_Transformer has a smaller error compared to the ORB_SLAM2, and the four error parameters are reduced by 18.450% 27.%, 18.177%, and 19.492%, respectively. Therefore, it is proved that the ORB_SLAM2_Transformer has an overall smaller positioning error and better relative and absolute positional errors. We believe that the ORB_SLAM2_Transformer achieves the goal of removing the dynamic feature point fraction to reduce the camera tracking localization error, thus optimizing the problem of camera tracking drift under dynamic targets.

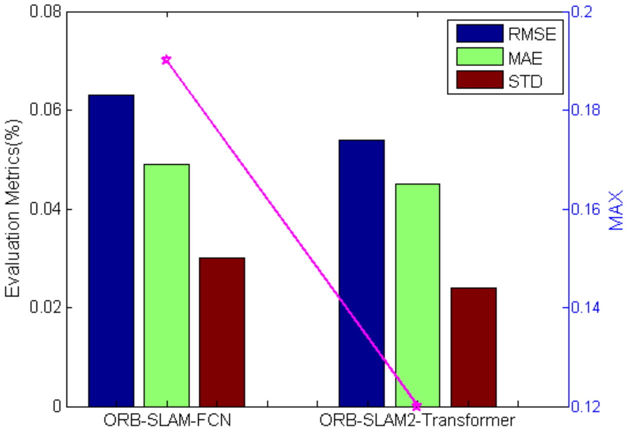


FIGURE 8. Main results about the APE performances of experimental methods on DataSet_Factory data set.

C. DISCUSSION

In this work, the deep learning-based image segmentation is employed to enhance the SLAM manufacturing process in terms of automobile parts. In our proposal, the Transformer is employed as backbone network for use. The performance of image segmentation methods directly determines efficiency of following additive manufacturing operations. Hence, the proposal is compared with a typical image segmentation method FCN for performance evaluation.

To better verify performance of the proposal, four aspects of evaluation metrics are introduced to visualize algorithm performance in the format of numerical values. The four aspects of metrics include: segmentation effect, time complexity, RPE performance, and APE performance. After simulative experiments on real-world scenes of SLAM-based additive manufacturing, the obtained results show that the proposal can have proper performance in terms of segmentation effect. The good image segmentation performance can well promote the following manufacturing operations.

Although the proposal can have proper performance in SLAM-based additive manufacturing process, there is still some distance to practical industrial application. The deep learning has received great development in recent years and brought much insight into many computer vision tasks. However, deep learning algorithms are mostly facing the problem of computational complexity, which requires relatively high hardware conditions [47]. Generalized into our proposal, how to improve the running efficiency and reduce computational complexity is the future direction of our work.

IV. CONCLUSION

In this paper, in order to achieve robust SLAM in dynamic scenes, a transformer based visual SLAM method is proposed. The method combines the segmentation technique with the motion-consistency detection algorithm. First, the transformer network is used to semantically segment the image to establish the a prior semantic information of dynamic objects, then the feature points belonging to dynamic objects are rejected by the motion-consistency detection algorithm. Finally, the static feature points are utilized for pose estimation and point cloud overlay. Simulation experiments are conducted to test function of the proposed method. The obtained results show that the absolute trajectory error and relative estimation error can be reduced additive manufacturing of vehicular parts compared with the traditional ORB_SLAM2 system.

REFERENCES

[1] R. Geyer, J. R. Jambeck, and K. L. Law, "Production, use, and fate of all plastics ever made," *Sci. Adv.*, vol. 3, no. 7, Jul. 2017, Art. no. e1700782.

[2] F. Wang, S. Fathizadan, F. Ju, K. Rowe, and N. Hofmann, "Print surface thermal modeling and layer time control for large-scale additive manufacturing," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 1, pp. 244–254, Jan. 2021.

[3] L. Chen, Y. Zhu, and C. K. Ahn, "Adaptive neural network-based observer design for switched systems with quantized measurements," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 10, 2021, doi: 10.1109/TNNLS.2021.3131412.

[4] Z. Guo, Y. Shen, S. Wan, W.-L. Shang, and K. Yu, "Hybrid intelligence-driven medical image recognition for remote patient diagnosis in Internet of Medical Things," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 12, pp. 5817–5828, Dec. 2022.

[5] A. Essien and C. Giannetti, "A deep learning model for smart manufacturing using convolutional LSTM neural network autoencoders," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 6069–6078, Sep. 2020.

[6] S. Liu, P. Gao, Y. Li, W. Fu, and W. Ding, "Multi-modal fusion network with complementarity and importance for emotion recognition," *Inf. Sci.*, vol. 619, pp. 679–694, Jan. 2023.

[7] S. Liu, S. Huang, S. Wang, K. Muhammad, P. Bellavista, and J. D. Ser, "Visual tracking in complex scenes: A location fusion mechanism based on the combination of multiple visual cognition flows," *Inf. Fusion*, Feb. 2023.

[8] Z. Zhou, X. Dong, Z. Li, K. Yu, C. Ding, and Y. Yang, "Spatio-temporal feature encoding for traffic accident detection in VANET environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19772–19781, Oct. 2022.

[9] P. Yu, C. Guo, Y. Liu, and H. Zhang, "Fusing semantic segmentation and object detection for visual SLAM in dynamic scenes," in *Proc. 27th ACM Symp. Virtual Reality Softw. Technol.*, Osaka, Japan, 2021, pp. 1–7.

[10] B. Zhu, K. Chi, J. Liu, K. Yu, and S. Mumtaz, "Efficient offloading for minimizing task computation delay of NOMA-based multiaccess edge computing," *IEEE Trans. Commun.*, vol. 70, no. 5, pp. 3186–3203, May 2022.

[11] Q. Zhang, K. Yu, Z. Guo, S. Garg, J. J. Rodrigues, M. M. Hassan, and M. Guizani, "Graph neural network-driven traffic forecasting for the connected Internet of Vehicles," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 5, pp. 3015–3027, Sep. 2022.

[12] M. Andronie, G. Lazaroiu, M. Iatagan, C. Uta, R. Stefanescu, and M. Cocosatu, "Artificial intelligence-based decision-making algorithms, Internet of Things sensing networks, and deep learning-assisted smart process management in cyber-physical production systems," *Electronics*, vol. 10, no. 20, p. 2497, Oct. 2021.

[13] G. Lazaroiu, M. Andronie, M. Iatagan, M. Geamanu, R. Stefanescu, and I. Dijmarescu, "Deep learning-assisted smart process planning, robotic wireless sensor networks, and geospatial big data management algorithms in the Internet of Manufacturing Things," *ISPRS Int. J. Geo-Inf.*, vol. 11, no. 5, p. 277, Apr. 2022.

[14] T. Kliestik, H. Musa, V. Machova, and L. Rice, "Remote sensing data fusion techniques, autonomous vehicle driving perception algorithms, and mobility simulation tools in smart transportation systems," *Contemp. Readings Law Social Justice*, vol. 14, no. 1, pp. 137–152, 2022.

[15] Y. Zhu and W. X. Zheng, "Observer-based control for cyber-physical systems with periodic DoS attacks via a cyclic switching strategy," *IEEE Trans. Autom. Control*, vol. 65, no. 8, pp. 3714–3721, Aug. 2020.

[16] S. Xia, Z. Yao, G. Wu, and Y. Li, "Distributed offloading for cooperative intelligent transportation under heterogeneous networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 16701–16714, Sep. 2022.

[17] J. Wei, Q. Zhu, Q. Li, L. Nie, Z. Shen, K. K. R. Choo, and K. Yu, "A redactable blockchain framework for secure federated learning in industrial Internet of Things," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17901–17911, Sep. 2022.

[18] Z. Guo, K. Yu, A. K. Bashir, D. Zhang, Y. D. Al-Otaibi, and M. Guizani, "Deep information fusion-driven POI scheduling for mobile social networks," *IEEE Netw.*, vol. 36, no. 4, pp. 210–216, Jul. 2022.

[19] S. Liu, Y. Li, and W. Fu, "Human-centered attention-aware networks for action recognition," *Int. J. Intell. Syst.*, vol. 37, no. 12, pp. 10968–10987, Dec. 2022.

[20] Y. Li, H. Ma, L. Wang, S. Mao, and G. Wang, "Optimized content caching and user association for edge computing in densely deployed heterogeneous networks," *IEEE Trans. Mobile Comput.*, vol. 21, no. 6, pp. 2130–2142, Jun. 2022.

[21] Y. He, L. Nie, T. Guo, K. Kaur, M. M. Hassan, and K. Yu, "A NOMA-enabled framework for relay deployment and network optimization in double-layer airborne access VANETs," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 22452–22466, Nov. 2022.

[22] R. Li, S. Wang, and D. Gu, "DeepSLAM: A robust monocular SLAM system with unsupervised deep learning," *IEEE Trans. Ind. Electron.*, vol. 68, no. 4, pp. 3577–3587, Apr. 2021.

[23] X. Zheng and Z. Cai, "Privacy-preserved data sharing towards multiple parties in industrial IoTs," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 968–979, May 2020.

[24] Z. Guo, K. Yu, Z. Lv, K.-K.-R. Choo, P. Shi, and J. J. Rodrigues, "Deep federated learning enhanced secure POI microservices for cyber-physical systems," *IEEE Wireless Commun.*, vol. 29, no. 2, pp. 22–29, Apr. 2022.

[25] D. Peng, D. He, Y. Li, and Z. Wang, "Integrating terrestrial and satellite multibeam systems toward 6G: Techniques and challenges for interference mitigation," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 24–31, Feb. 2022.

[26] Z. Cai and Q. Chen, "Latency-and-coverage aware data aggregation scheduling for multihop battery-free wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1770–1784, Mar. 2021.

[27] Z. Guo, K. Yu, A. Jolfaei, F. Ding, and N. Zhang, "Fuz-Spam: Label smoothing-based fuzzy detection of spammers in Internet of Things," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 11, pp. 4543–4554, Nov. 2022.

[28] L. Zhao, H. Chai, Y. Han, K. Yu, and S. Mumtaz, "A collaborative V2X data correction method for road safety," *IEEE Trans. Rel.*, vol. 71, no. 2, pp. 951–962, Jun. 2022.

[29] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyber-physical systems," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 2, pp. 766–775, Apr. 2020.

[30] Z. Xing, X. Zhu, and D. Dong, "DE-SLAM: SLAM for highly dynamic environment," *J. Field Robot.*, vol. 39, no. 5, pp. 528–542, Aug. 2022.

[31] B. Fang, X. Han, Z. Wang, and X. Yuan, "SLAM algorithm based on bounding box and deep continuity in dynamic scene," *Int. J. Wireless Mob. Comput.*, vol. 21, no. 4, pp. 349–364, 2021.

[32] J. Chang, N. Dong, and D. Li, "A real-time dynamic object segmentation framework for SLAM system in dynamic scenes," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[33] J. Ni, X. Wang, T. Gong, and Y. Xie, "An improved adaptive ORB-SLAM method for monocular vision robot under dynamic environments," *Int. J. Mach. Learn. Cybern.*, vol. 13, no. 12, pp. 3821–3836, Dec. 2022.

[34] Y. Fan, Q. Zhang, S. Liu, Y. Tang, X. Jing, J. Yao, and H. Han, "Semantic SLAM with more accurate point cloud map in dynamic environments," *IEEE Access*, vol. 8, pp. 112237–112252, 2020.

[35] C. Shao, C. Zhang, Z. Fang, and G. Yang, "A deep learning-based semantic filter for RANSAC-based fundamental matrix calculation and the ORB-SLAM system," *IEEE Access*, vol. 8, pp. 3212–3223, 2020.

[36] N. Ragot, R. Khemmar, A. Pokala, R. Rossi, and J.-Y. Ertaud, "Benchmark of visual SLAM algorithms: ORB-SLAM2 vs RTAB-map," in *Proc. 8th Int. Conf. Emerg. Secur. Technol. (EST)*, Colchester, U.K., Jul. 2019, pp. 1–6.

[37] L. Cui and C. Ma, "SDF-SLAM: Semantic depth filter SLAM for dynamic environments," *IEEE Access*, vol. 8, pp. 95301–95311, 2020.

[38] I. A. Kazerouni, L. Fitzgerald, G. Dooly, and D. Toal, "A survey of state-of-the-art on visual SLAM," *Exp. Syst. Appl.*, vol. 205, Nov. 2022, Art. no. 117734.

[39] Z. Cai, Z. Duan, and W. Li, "Exploiting multi-dimensional task diversity in distributed auctions for mobile crowdsensing," *IEEE Trans. Mobile Comput.*, vol. 20, no. 8, pp. 2576–2591, Aug. 2021.

[40] D. Feng, Z. Zhang, and K. Yan, "A semantic segmentation method for remote sensing images based on the Swin transformer fusion Gabor filter," *IEEE Access*, vol. 10, pp. 77432–77451, 2022.

[41] M. Dai, J. Hu, J. Zhuang, and E. Zheng, "A transformer-based feature segmentation and region alignment method for UAV-view geo-localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4376–4389, Jul. 2022.

[42] Y. Yang, X. Zhao, M. Huang, X. Wang, and Q. Zhu, "Multispectral image based germination detection of potato by using supervised multiple threshold segmentation model and Canny edge detector," *Comput. Electron. Agricult.*, vol. 182, Mar. 2021, Art. no. 106041.

[43] J. Zhao, L. Zhao, S. Huang, and Y. Wang, "2D laser SLAM with general features represented by implicit functions," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4329–4336, Jul. 2020.

[44] A. Bojko, R. Dupont, M. Tamaazousti, and H. L. Borgne, "Learning to segment dynamic objects using SLAM outliers," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Milan, Italy, Jan. 2021, pp. 9780–9787.

[45] A. Handa, T. Whelan, J. McDonald, and A. J. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, Hong Kong, 2014, pp. 1524–1531.

[46] K. Yang, Y. Liu, S. Zhang, and J. Cao, "Surface defect detection of heat sink based on lightweight fully convolutional network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[47] Z. Cai, Z. He, X. Guan, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Trans. Depend. Sec. Comput.*, vol. 15, no. 4, pp. 577–590, Aug. 2018.

WENBO XU received the B.S. degree from Henan University of Technology, in 2015, and the M.S. degree from Xi'an University of Science and Technology, in 2018. He is currently a Lecturer with the School of Vehicle and Traffic Engineering, Henan Institute of Technology. His research interests include the application of additive manufacturing in industry and the design of 3-D vehicle modeling.

WEIWEI FAN received the Ph.D. degree from Dalian University of Technology. He is currently an Associate Professor with Henan Institute of Technology. He is mainly engaged in the research of spray combustion in engine and numerical simulation of chemical reaction kinetics.

JINGYANG LI received the B.S. and M.S. degrees from Shenyang Agricultural University, in 2006 and 2009, respectively. He is currently a Lecturer with the School of Mechanical Engineering, Henan Institute of Technology. His research interests include the application of additive manufacturing in industry and the design of hydraulic and pneumatic transmission systems.

OSAMA ALFARRAJ received the master's and Ph.D. degrees in information and communication technology from Griffith University, in 2008 and 2013, respectively. He is currently a Professor of computer science with King Saud University, Riyadh, Saudi Arabia. His current research interests include eSystems (eGov, eHealth, and ecommerce), cloud computing, and big data. He was a Consultant and a member of Saudi National Team for Measuring E-Government, Saudi Arabia, for two years.

AMR TOLBA (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees from the Department of the Mathematics and Computer Science, Faculty of Science, Menoufia University, Egypt, in 2002 and 2006, respectively. He is currently a Full Professor of computer science with King Saud University (KSU), Saudi Arabia. He has authored/coauthored over 130 scientific articles in top-ranked (ISI) international journals, such as IEEE Internet of Things Journal (IoT), *ACM Transactions on Internet Technology* (TOIT), *IEEE Consumer Electronics Magazine* (CEMAG), IEEE Access, IEEE Systems Journal, *Future Generation Computer Systems* (FGCS), *Journal of Network and Computer Applications* (JNCA), *Neural Computing and Applications* (NC&A), *Journal of Ambient Intelligence and Humanized Computing* (JAIHC), *Computer Networks* (COMNET), *Computer Communications* (COMCOM), P2PNET, VCOM, and *World Wide Web Journal* (WWWJ). He has translated four books into the Arabic language. His main research interests include artificial intelligence (AI), the Internet of Things (IoT), data science, and cloud computing. He has served as a Technical Program Committee (TPC) Member at several conferences, such as DSIT 2022, CICA 2022, EAI MobiHealth 2021, DSS 2021, AEMCSE 2021, ICBDM 2021, ICISE 2021, DSS 2020, NCO 2020, ICISE 2019, ICCSEA 2019, DSS 2019, FCES 19, ICISE 2018, ESG!/18, Smart Data!/17, NECO 2017, NC!/17, WEMNET!/17, NET!/17, and Smart Data!/16. He has been included in the list of the top 2% of influential researchers globally (prepared by scientists from Stanford University, USA) in 2020, 2021, and 2022. He served as an associate editor/guest editor for several ISI journals.

TIANHONG HUANG received the B.S. degree from East China Jiaotong University, Nanchang, China. He is currently pursuing the M.S. degree in computer science with Oregon State University, Corvallis, OR, USA. His research interests mainly include quantization, natural language processing, machine learning, and deep learning.