ARCargo: Multi-Device Integrated Cargo Loading Management System with Augmented Reality

1st Tianxiang Zhang*
Beijing Institute of Spacecraft System Engineering
braveztx@163.com

2nd Chong Bao*
State Key Lab of CAD&CG
Zhejiang University
chongbao@zju.edu.cn

3rd Hongjia Zhai*
State Key Lab of CAD&CG
Zhejiang University
zhj1999@zju.edu.cn

4th Jiazhen Xia
State Key Lab of CAD&CG
Zhejiang University
xiajiazhen@zju.edu.cn

5th Weicai Ye
State Key Lab of CAD&CG
Zhejiang University
weicaiye@zju.edu.cn

6th Guofeng Zhang†
State Key Lab of CAD&CG
Zhejiang University
zhangguofeng@zju.edu.cn

***Abstract*—Traditional cargo loading requires a static recording of cargo information and status on paper. Storekeepers need to arrange cargoes by memory or cargo menus, severely restricting the entire cargo loading management process. To address this problem, we present a multi-device integrated cargo loading management system with Augmented Reality (AR), termed ARCargo, which monitors cargoes by fusing perceptual information from multiple devices in real-time. Then, we propose a visual localization method using hybrid features to strengthen the localization accuracy. For providing an intuitive and user-friendly interactive way, we design AR-driven guidance and monitoring alarms to reduce the workload of cargo loading management. Extensive experiments show that our system can efficiently, intelligently, and conveniently carry out cargo loading management operations.**

***Index Terms*—Cargo Loading Management System, Multi-Device Integrated, Augmented Reality, Visual Localization, SLAM, AprilTag**

I. Introduction

A timely and convenient cargo loading management system is essential to ensure a good experience for employees in cargo loading and achieve better cost reduction and efficiency. Recently, augmented reality (AR) technology [1] has gained wide attention, with applications ranging from entertainment, live broadcast, navigation, etc. However, it has rarely been used in the area of cargo loading management. Traditional cargo loading management systems involve considerable administrative work, including cargo information management, cargo scheduling, monitoring, adjustment on cargo status. A high level of accuracy, convenience, and real-time is urged in cargo loading. Universally, cargo information, such as material, weight, size, and location, is recorded statically on paper, which costs dozens of paper documents. In traditional cargo loading operations, the storekeeper needs to place the cargo in the designated position by memory or cargo menu, which aggravates their physical and psychological burdens and even leads to an inaccurate placement that severely limits the efficiency of the overall cargo loading. Throughout the loading process, the position or status of the cargoes may change unexpectedly (for example, cargoes may fall from the shelves due to collision), which requires storekeepers to spend extra time on risk detection and manual correction. For this reason, we consider the possibility of introducing AR technology to enhance the cargo loading management system, as illustrated in Fig. 1.

Fig. 1. **ARCargo system.** Our system supports various mobile devices with AR, such as phones, tablets and head-mounted display (HMD). (a) Our system works with a smartphone HUAWEI P20 Pro. (b) Our system works with a AR glass Shadow Creator Action One Pro.

For the problems aforementioned, we do the following analysis: AR-supported cargo records can render the cargo information onto cargo entities, enabling the integration of virtuality and reality. With AR devices, storekeepers can instantly get information about which item to pick and where to place it without searching dozens of paper documents. With the AR interactive guidance, the cargo arrangement becomes traceable and more convenient. To ensure the correct placement of cargo, a multi-device collaborative cargo loading system is required to monitor cargo locations in real-time and alert the storekeeper for unexpected cargo changes. AR can provide continuous cargo information in the dynamic loading process, which can accelerate the whole procedure. In addition, the system is required to support multiple people to perform loading management operations simultaneously.

* Equal contribution.
† Corresponding author.

Based on the above considerations, we propose an integrated multi-device cargo loading management system with AR. Our system, illustrated in Fig. 3, includes a cloud-based server to manage cargo information and a 3D model of the scene and perform cargo distribution, visual localization, monitoring, and other tasks. It allows different mobile devices, such as smartphones, tablets, or HMD, to communicate with each other to perform tasks such as cargo information management, interactive cargo guidance, etc. Please see the method section III for more details. The monitoring module of the system is ensured by a stereo camera and the clients, where the stereo camera is statically placed to provide the global information of the shelves, and the clients dynamically track the cargo's local information, which greatly enhances the robustness of the system. Since the shelves are usually covered by weak texture and repetitive structure, the accuracy of traditional visual localization methods is significantly reduced. To solve this problem, we propose a visual localization algorithm that combines markers features with scene features. The direct use of detected markers and scene features may not improve the accuracy because the accuracy of markers closer to the visual frustum surpasses the scene features and vice versa. Therefore, we propose a hybrid feature-based visual localization algorithm weighted by uncertainty.

In terms of reducing the burden of workers on cargo loading management, we propose a new AR-support interaction mode, shown in Fig. 2, which can inform the cargo details, the target location, as well as an arrow to guide the workers to perform the relevant operations. As for monitoring and alarm, our system's hybrid visual localization module will alert workers if some cargoes are misplaced accidentally so that the storekeeper can revise them immediately. Various experiments have proven the accuracy of our system, and user studies with objective and subjective criteria have shown that our AR-enabled system can facilitate intelligent and efficient loading operations and further ease the mental and physical burden of workers. Overall, our contributions are summarised as four-fold.

- We propose a novel multi-device integrated cargo management system with AR.
- We present a visual localization method using hybrid features based on uncertainty.
- We introduce an AR-driven interactive guidance and monitoring alarm program to improve the cargo loading management experience.
- Extensive experiments have shown that our system can efficiently, intelligently and conveniently carry out cargo loading operations.

II. Related Work

A. Simultaneous Localization and Mapping

Simultaneous localization and mapping (SLAM) mean using images to recover 3D structures and sensor motions in unknown environments [1]. It was initially proposed to equip autonomous robots with various sensors to locate their position and pose. In recent years, SLAM techniques that use the camera only have been widely discussed. Since the camera only can capture images, this type of SLAM is called visual SLAM (VSLAM). VSLAM can provide accurate pose estimation for mobile devices, allowing AR applications to realistically superimpose virtual objects over real-world [1]. The first solution to the monocular VSLAM is MonoSLAM [2], which is based on the extended Kalman filter model. Subsequently, the keyframe strategy is introduced to SLAM. ORBSLAM [3] extracts ORB features, builds covisibility graph to reduce the complexity of tracking and mapping and enables loop closure with DBoW2 [4] for building a complete and accurate global map. Recently, a multi-sensor fusion strategy has been applied to enhance the robustness of VSLAM under extreme situations. The first visual odometry system with tightly coupled inertial measurement units (IMUs) is OKVIS [5], which fuses motion measurement of IMU in visual odometry. However, the robustness of OKVIS is severely limited by the speed of IMU initialization. VINS-Mono [6] is a successful monocular visual-inertial odometry with pose graph optimization, map-merging, loop closure with DBoW2 and feature tracking with Lucas-Kanade tracker [7]. In our system, the phones and tablets use monocular visual-inertial SLAM for tracking, while the AR glasses use stereo visual-inertial SLAM.

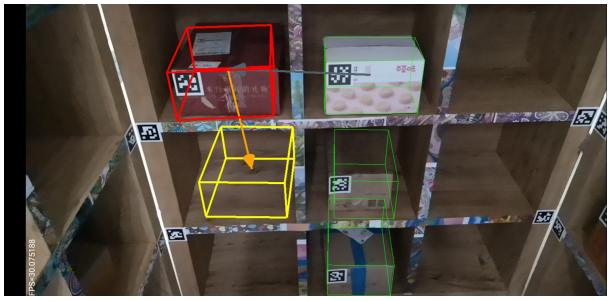

Fig. 2. **Visualization of our AR guidance.** The red cube denotes the current location of the cargo. The yellow cube denotes the target location of the cargo. The green cube denotes that the cargo was correctly placed.

B. Visual Localization

Accurate visual localization pays attention to find the correspondences between 2D key points of the query image and 3D points in a 3D SfM model. Those methods often follow the pipeline, detect key points, extract the description of key points and perform matching between 2D and 3D points to estimate the pose of the given query image. To scale the localization model to large scenes [8]–[11], they leverage image retrieval based method [12], [13] to filter dissimilar candidates. Using retrieval-based method only can provide a coarse pose for localization [14]. Besides, approaches that directly regress a single image pose are not competitive in terms of accuracy. HLoc [10] is the state-of-the-art method for visual localization task, which combines the superior visual place recognition approach, NetVLAD [15], 2D keypoint detector, SuperPoint [16] and feature matching method SuperGlue [17]. HLoc can achieve good performance under strong appearance

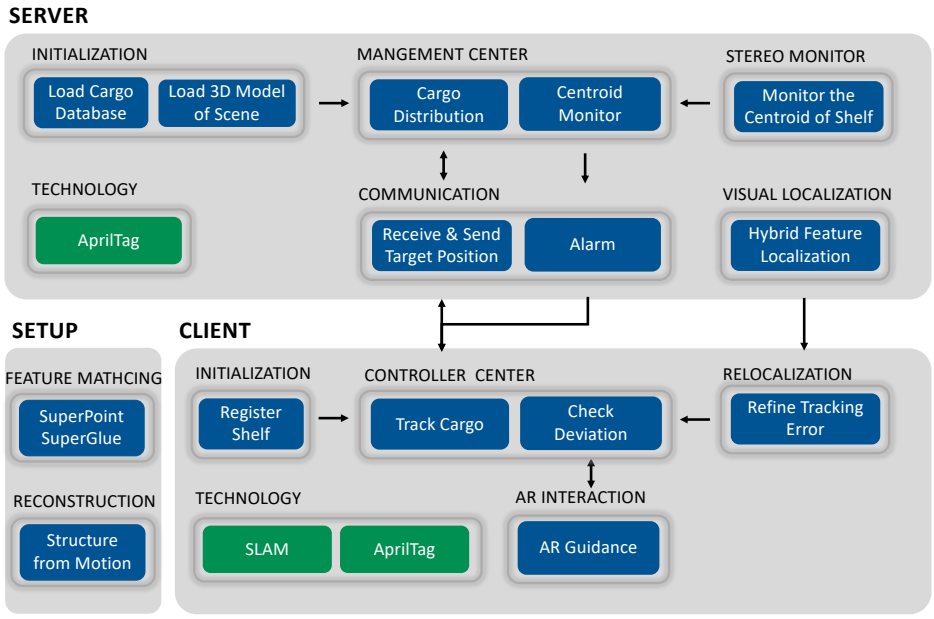

Fig. 3. **ARCargo architecture.** Our architecture consists of a cloud-based server and multiple mobile clients. The server is responsible for the storage management of cargoes, which provides services such as cargo distribution, centroid monitoring and communication with the client about the location of cargoes. While the clients rely on SLAM, AprilTag tracks user movements and real-world objects for providing accurate AR assistance in cargo handling.

changes and different light conditions thanks to those previous works. Since our shelf scenes are challenging and contain many textureless structures, the state-of-the-art methods such as HLoc [10] are difficult to always perform robust pose estimation, and we additionally add AprilTag [18] markers. Considering the difference in the confidence level of different features, we propose an uncertain-aware visual localization method using hybrid features.

C. AR Systems

AR refers to the real-time technology that enhances real-world 3D environments with virtual 3D objects to aid the user's interaction with the real world [1]. AR systems have been integrated into all kinds of commercial mobile devices, thanks to the development of SLAM, visual localization, 3D rendering, and AR Software Development Kits (SDK) that package these underlying technologies as easy-to-use interfaces. Among the most widely used AR SDKs are ARKit for iOS, ARCore for Android, and Vuforia for the Unity game engine [19]. Meanwhile, the fast development of AR systems has also been spurred by advances in computational power, and various hardware components such as a head-mounted display (HMD), positional tracking sensors, gestural input devices, and wireless communication components [20]. These hardware developments allow us to get rid of large cumbersome proprietary AR devices and embrace commercially affordable portable devices, such as Microsoft HoloLens, that facilitate hands-free user interactions and immersive experiences [21]. While we develop different mobile AR programs according to the hardware properties of different mobile devices and cooperate with our cloud server architecture to realize a multi-device integrated cargo loading modelling management system with AR.

III. Multi-Device Integrated Cargo Management System with AR

A. Multi-Device Integrated Cargo Management System

The architecture of our system, i.e. *ARCargo*, works as a client-server model, shown in Fig. 3. The server is responsible for the storage management of cargoes, which provides services such as cargo distribution, centroid monitoring and communication with the client about the location of cargoes. Furthermore, we deploy an additional stereo vision camera to monitor cargoes. The clients rely on SLAM, AprilTag to track user movements and real-world objects for providing accurate AR assistance in cargo handling. The video showing the use of the system is available at https://youtu.be/La7TNMIDWvY.

1) Server: The server is deployed in the cloud, and we have also designed a desktop application to provide visual feedback and operational experience. First, the server will automatically load the cargo database and the 3D model of shelves, which is built by COLMAP [22], [23] in advance. After receiving cargo loading requests from the client, the server will send the target cargo location to the client. The target location is derived from our assignment algorithm, which is designed to enhance the stability of shelves. The instability is equivalent to the deviation of the shelves centroid caused by the placement of the cargo. The client will report the final location of cargo to the server after cargo is well placed, and the server checks its centroid deviation to the target location. Since the wrong placement, which significantly deviates the overall centroid from the stable location, may lead to the shelf's corruption or

damage, the server's alarm module will send a signal to the client and ask it to place the cargoes at the correct position.

In addition, we equip a stereo camera to monitor the distribution of cargoes and the centroid of cargoes shelves in real-time. We put AprilTag makers on the shelf sidebar, shown in Fig. 4, and obtain the shelf coordinate system from the 3D model. AprilTag markers are used to establish the correspondences between the camera and the shelf coordinate system. PnP [24] is employed to efficiently solve the relative transformation between the camera and shelf coordinate system from which the centroid of shelf results. The cargo is also tagged with AprilTag markers so that each cargo can be identified uniquely, and the stereo camera can track its centroid. The server checks the deviation of the overall centroid by fusing the centroids of cargoes and shelves captured by the stereo camera statically. On the other hand, the server also dynamically utilizes the information from the client to track the centroid of cargoes where the stereo camera can not see. The multi-device integrated centroid monitoring approach ensures the robustness and accuracy of the system.

Furthermore, a hybrid feature-based visual localization module is developed to improve the quality of pose estimation at the client. We extract the sparse feature points and corner points of AprilTag from the received image and fuse them to reach a more accurate localization result. We detail it in subsection III-B.

Fig. 4. The cargo shelves tagged with AprilTag markers.

2) Clients: Our clients can adapt to various devices, such as HMDs and mobile phones. The clients exploit SLAM to track the position and orientation of the device. In the initialization phase, SLAM is activated first to track the device and build the initial map. With AprilTag markers on the shelf, we can easily align the SLAM coordinate system to the shelf coordinate system by solving the PnP problem, depicted in Fig. 5. When an AprilTag marker (attached to the cargo) with high confidence is recognized, the client asks the remote server for the target location of the cargo. We set translation and rotation constraints to determine the deviation of the centroid of the cargo. As soon as the deviation of the centroid exceeds the threshold, the client activates the AR guidance module to guide the user to place the cargo at the correct location. After the cargo is well placed, the client sends its final location to the server, which helps the server monitor the cargoes dynamically.

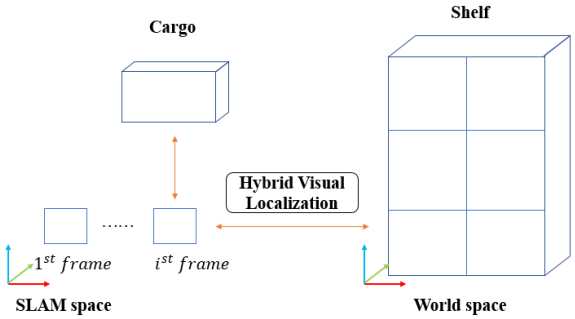

Fig. 5. **The coordinate transformation of ARCargo schematic overview.** We define the coordinate system of reconstruction of shelves as the world space. SLAM tracks the pose of continuous frames. Hybrid visual location is used to compute the transformation between each frame and shelf. The cargo is registered to SLAM space by AprilTag.

The pose estimation based on AprilTag is severely affected by the distance between the camera and AprilTag. When the distance is too large, the corner points of AprilTag account for a small part of the image, which exacerbates the uncertainty of pose estimation. Visual localization is a well-defined solution for indoor localization at an arbitrary scale, which we apply to resolve the pose estimation at a great distance. When the client detects that the distance between the device and AprilTag is too large, it sends the current frame to the server. The server performs visual localization with hybrid features and responds to the client with the solved relative poses (about 1FPS), allowing the client to initialize more robustly. Moreover, after the client has been running for a long time (about every 100 frames), the client will request the global localization results from the server to correct the drift, which ensures our system can run in real time and robustly.

B. Visual Localization using Hybrid Features

The localization module will fail if we only use ApriTag markers to compute the pose of the given query image when the image does not contain AprilTag markers or is far away from the shelves. Inspired by the visual localization technique, we combine the visual localization method and the AprilTag-based method to get the pose of the current image robustly. The pipeline is shown in Fig. 6.

1) HLoc: Given a query image I_q captured by the mobile device, the localization module will return the pose of the given query image. To achieve this goal, we obtain the pose in the following steps. Firstly, different from HLoc [10], we use the NetVLAD model to retrieve the similar scene in the database, which are collected in the system setup phase. NetVLAD is more powerful in visual place recognition than the MobileNet [25] used in HLoc:

$$D_q = F_{NetVLAD}(I_q) \tag{1}$$

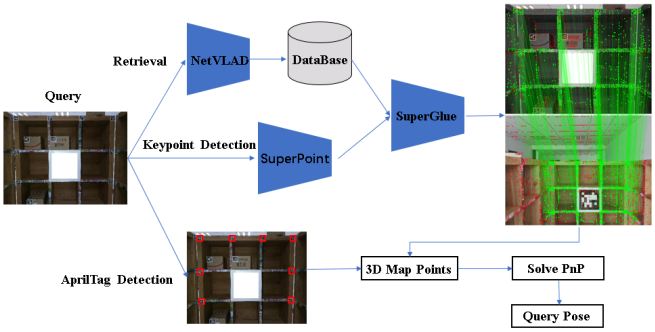

Fig. 6. **Pipeline of Visual Localization Using Hybrid Features.** We perform hierarchical visual localization based on the uncertainty of the hybrid features of marker and scene.

The output of NetVLAD is a global descriptor of the image which will be used to search the similar descriptor in the database:

$$D_{db} = \{F_{NetVLAD}(I_i)|i \in [1, N]\} \tag{2}$$

where $F_{NetVLAD}$ represents the convolutional and fully connected layers of the NetVLAD model. N is the size of the image database, $I_q, I_i, i \in [1, N]$ are the query and database images, respectively. D_q, D_{db} are the keypoint descriptors of query and database images.

After obtaining the global descriptors of query and database images, we can get Top-K database images that are similar to the query image by using the nearest neighbour search method:

$$I_{Topk} = NN_k(D_q, D_{db}) \tag{3}$$

We detect the 2D keypoints of the query and database images using the SuperPoint model. K_{Lq}, K_{Dq} are the locations, descriptions of the 2D keypoint for the query images, respectively:

$$K_{Lq}, K_{Dq} = F_{SuperPoint}(I_q) \tag{4}$$

K_{Ldb}, K_{Ddb} are for the database images similarly:

$$K_{Ldb}, K_{Ddb} = \{F_{SuperPoint}(I_i)|i \in [1, N]\} \tag{5}$$

Now we have obtained the Top-k retrieval results and the locations and descriptors for 2D keypoints. We use SuperGlue to match features and get 2D-2D correspondences. With the 3D SfM model, we can obtain the 3D locations of the key points in the database images. So, we can get the 2D-3D correspondences and solve the PnP problem to get the camera pose of the query image. For more details about those convolutional neural networks we used in the part, please refer to NetVLAD [15], SuperPoint [16] and SuperGlue [17].

2) AprilTag: When closing to the shelf or AprilTag, using the AprilTag for localization will get a good performance. AprilTag is a visual fiducial system, useful for various tasks, including AR, robotics, and camera calibration. Targets can be created from an ordinary printer, and the AprilTag detection software computes the precise 3D position, orientation, and identity of the AprilTag markers relative to the camera [26]. However, in those places, the camera is far away from the AprilTag, the pose solved from AprilTag marker is inaccurate. So using HLoc can get a good localization performance. The 3D positions of 2D keypoints are already estimated from multi-view images in the SfM stage. So, we can accurately estimate the rotation matrix of the camera using the obtained 2D-3D correspondences. We use HLoc to obtain a relatively accurate pose from the place where the camera can not see the AprilTag markers.

3) Combined Visual Localization: The shelf system has many repetitive structures and weak textures. If we only use the 2D-3D correspondences from visual localization, the localization module will fail sometimes. So we combine the AprilTag corner detection with visual localization. When we stand far away from the AprilTag, the AprialTag corners have high uncertainty and cannot be recognized. So, it is hard to know where we are. In this situation, we can use the appearance and structure information for the localization task. We use SuperPoint [16] and SuperGlue [17] to detect 2D keypoints and match features to get the correspondences. When we are near the shelf, we can see many repetitive grids, and the visual features extracted from convolutional neural networks are similar to other locations. Under the situation, the accuracy of visual location sharply decreases. So we use the AprialTag information for localization. Q_{tag}, T_{tag} is the pose (rotation and translation) solved from AprilTag and Q_{vl}, T_{vl} is the pose from visual localization. Here we use quaternion to represent the rotation. To combine those two poses, we obtain the interpolation of those poses. The weight between the two poses depends on the size of the AprilTag that we captured in the image.

$$Q_{hybrid} = slerp(Q_{tag}, Q_{vl}, \alpha) \tag{6}$$

where $slerp()$ is the interpolation function for quarternion.

$$T_{hybrid} = \alpha \times T_{tag} + (1 - \alpha) \times T_{vl} \tag{7}$$

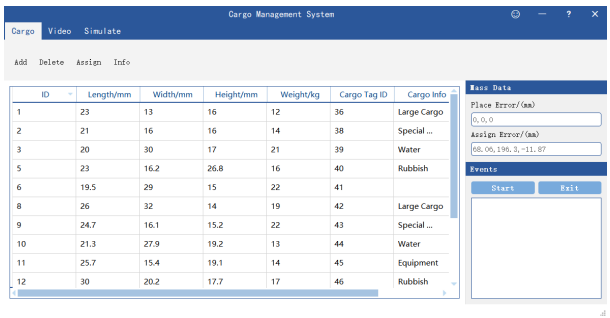

Fig. 7. The UI of ARCargo Server.

C. An AR Interactive Pipeline

1) Cargo Information Management: The Windows application running on the server supports various administrative operations ranging from cargo information entry to cargo distribution optimization, which is performed in a visualized

way through a 2D shelf model displayed on the PC screen, as shown in Fig. 7. Moreover, warning information is given on the Windows PC when the centroid of the cargoes exceed safe limits.

2) Interactive Guidance with AR: Real-time instructions and hints are given to the storekeeper through virtual 3D shapes displayed on the Android phone and AR glasses carried by the storekeeper during the entire work pipeline, shown in Fig. 8. The storekeeper starts the work pipeline by scanning several AprilTag markers on the shelves with a mobile AR device so that the device can be initialized. When the device has encountered enough AprilTag markers (usually 5 to 8) to infer its initial location, a visual cue is given to the storekeeper to proceed with their work. Next, the mobile device continuously captures all visible AprilTag markers. For each cargo box lying within the current field of view, the ID of the box is decoded from the AprilTag. A query will be sent to the server to get the destination location of this box. Meanwhile, the current location of the box is calculated from the relative pose of AprilTag along with the device's pose informed by SLAM, show in Fig. 5. All those cargo boxes whose current location is more than 3 centimetres away from the destination are considered in need of manual transportation and are added to the transportation queue. If the queue is empty, all cargo boxes are highlighted with green virtual boxes to indicate the completion of work; otherwise, the cargo box at the front of the transportation queue is highlighted with a virtual red box, with a virtual arrow pointing from the centroid of its current location to the centroid of destination.

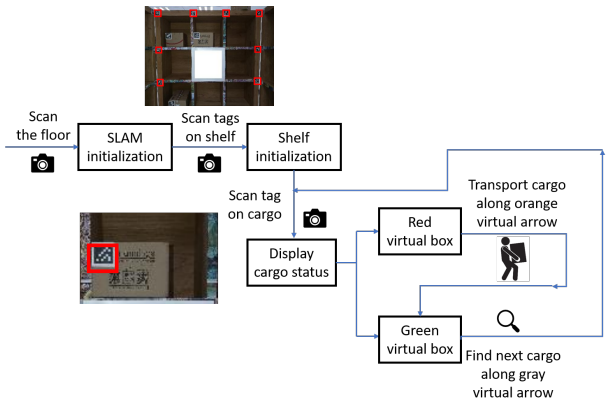

Fig. 8. **Pipeline of Interactive Guidance.**

3) Monitoring and Alarm: When each cargo have been placed in its assigned location, the PC system will monitor the movement of every cargo. Once the cargo has been moved bigger than a threshold, the PC system will send a message to the mobile device and print some information to the screen, which will warn the people to pay attention to those cargoes. Then an AR interactive guided cargoes adjustment will be launched.

IV. Experiments

A. System Setup

We need to build a 3D map of the shelves. Firstly, to improve the discrimination of different parts of the shelves, we add some AprilTag markers and colourful pictures on the shelves. Secondly, we use a handheld camera to capture the texture information of the shelf at different viewpoints and locations. Then, we use HLoc [10] to build the sparse 3D SfM model, which uses SuperPoint [16] to detect and describe the 2D keypoints and SuperGlue [17] to match keypoints. Besides, we use the pose recovered from the last step to triangulate the AprilTag corners and obtain the 3D locations in the sparse model. Finally, we use the 3D point cloud for visual localization. The sparse 3D SfM model is shown in Fig. 9, which contains the recovered camera poses and 3D points. The SLAM function is supported by the AR device itself with corresponding SDK. For instance, we can use HUAWEI AR Engine for HUAWEI P20 pro.

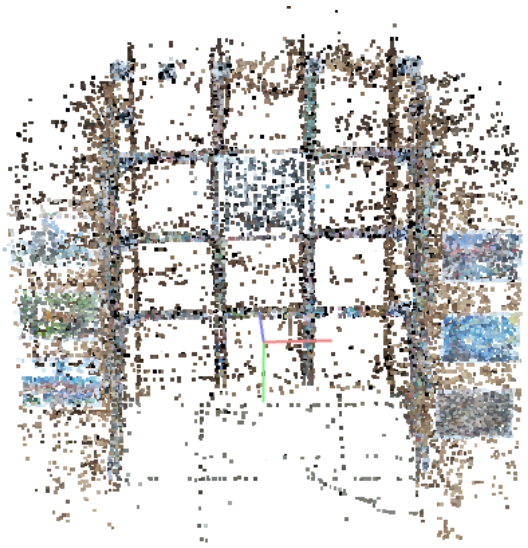

Fig. 9. **Sparse 3D SfM model of the shelves.**

B. Hybrid Visual Localization Results.

The visual localization task is to estimate the camera pose of an image given the image and its 3D geometric model. In a realistic scenario of a cargo loading modelling system, it is difficult to obtain a ground-truth camera pose. For this reason, we perform the pose calibration by posting a gaint marker as the landmark to the scene. To obtain the reconstructed shelf model, we reconstruct the scene geometry with the gaint marker using COLMAP [22], [23] and obtain the registered poses of each image. In the visual localization evaluation, we capture hundreds of frames containing the gaint marker evenly in the scene and use them to solve the camera pose as the ground-truth poses. To measure the effectiveness of our visual localization method, we erase the large markers from each given image and only use the small markers, scene features and hybrid features for visual localization. Fig. 10

TABLE I

ABLATION STUDY OF HYBRID LOCALIZATION. WE REPORT THE POSE RECALL [%] AT DIFFERENT ACCURACY LEVELS OF POSITIONS (m) AND ORIENTATIONS (deg) ON OUR ARCARGO DATASETS. THE ARCARGO DATASETS WERE TAKEN AT DIFFERENT DISTANCES FROM THE SHELF. HYBRID LOCALIZATION ACHIEVE A GOOD BALANCE AND BE MORE ROBUST TO THE CHALLENGES OF NEAR AND FAR SCENES.

	ARCargo Dataset			
	orientation [deg]	distance[m]	average	average
method	1.5, 2.0, 2.5	0.03, 0.04, 0.05	orientation[deg]	distance[m]
only marker	96.3 / 96.3 / 96.3	74.1 / 88.9 / 88.9	4.98	0.16
only HLoc	77.8 / 88.9 / 96.3	63.0 / 70.4 / 74.1	1.14	0.26
marker+HLoc(hybrid)	85.2 / 92.6 / 96.3	63.0 / 66.7 / 70.4	2.98	0.20

shows the matching results obtained from the given image and the adjacent frames retrieved from the image, and it can be seen that many matching points can be obtained.

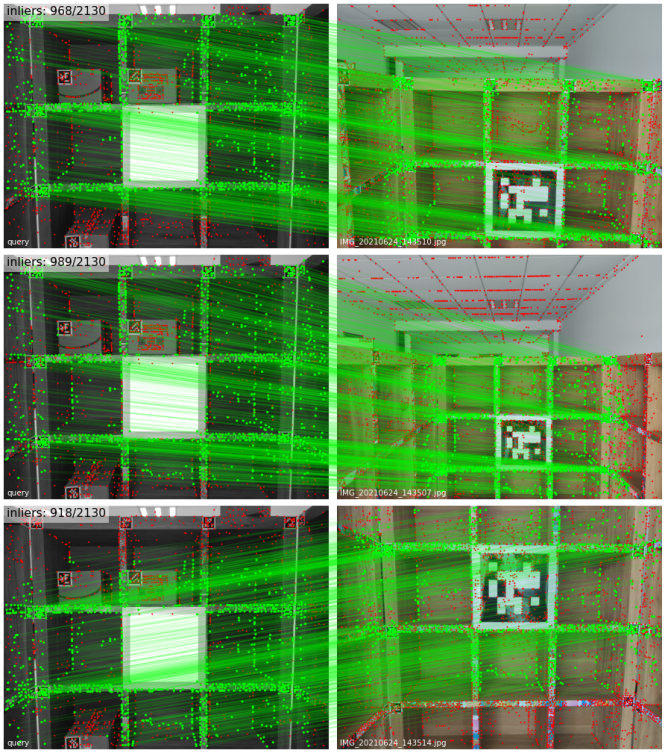

Fig. 10. **Mathcing results**. The left column contains the query images on each row, and the right column contains the retrieved nearest database images. All the correctly matched 3D points are projected onto this database image, and the projections beyond the border are not displayed.

1) Ablation Study: In this part, we show the performance of different visual localization results on the ARCargo datasets, taken at different distances from the shelf. The number of images taken closer to the shelf is much greater than the number of pictures taken farther away from the shelf. From Table I, we can see that the pose estimation using the only marker is better than the hierarchical visual localization results at different bins. However, the average rotation error is very large. The AprilTag accuracy is good enough to obtain a higher percentage when the client is close to the shelf. However, when it is far away from the shelf, the quality of AprilTag detection will be significantly weakened, and the error will be very large, making the overall error large. As for the hierarchical visual localization method, since it only considers the points of the scene, in our repeated-texture or weak-texture shelf geometry model, it is easy to have too many false matches when the client is close to the shelf. So, our hybrid visual localization algorithm considers the characteristics of the two different features and adapts according to uncertainty information such as distance to the shelf, which can achieve a good balance and be more robust to the challenges of near and far scenes.

C. User study

12 volunteers (8 males and 4 females) were recruited to take part in the study. Firstly, participants were asked to place cargoes at the correct position on the shelves, helped by a 2D distribution map. Secondly, we assigned AR devices to each user (6 persons use HUAWEI P20 Pro and 6 others use Shadow Creator Action One Pro). Participants were requested to perform another cargo loading guided by AR promoting. In this study, we captured the objective, as well as subjective, criteria. We physically measured the system's error by calculating the centroid deviation of the cargoes from the target position resulted from the server. After each transport, a questionnaire about the interaction was completed by participants at once.

TABLE II

ACCURACY OF CARGO BOX PLACEMENT. OUR AR-BASED INTERACTIVE GUIDANCE METHOD IS MORE EFFECTIVE IN ENHANCING CARGO PLACEMENT ACCURACY THAN THE METHOD WITHOUT GUIDANCE.

Deviation of Centroid (cm)	**Cargo Assignments**			
	#1	***#2***	***#3***	**Average**
without guidance	4.9	5.2	4.1	4.7
guided by ARCargo (ours)	**1.3**	**1.5**	**1.8**	**1.5**

1) Assistance and Accuracy: Given three different assignments of cargo distribution, We measure the deviation of the centroid of eight cargo boxes from the target position, with or without the guidance of ARCargo. The results are shown in Table II, indicating that our AR-based interactive guidance method is effective in enhancing cargo placement accuracy.

2) Interaction: We utilized NASA-TLX questionnaire [27] to measure the subjective workload of our AR interaction for cargo loading. After the operation, participants were asked to score in six dimensions from 5 to 100 (mental demand, physical demand, temporal demand, performance, effort, and frustration), where the lower score denotes the participant feels ease, relaxed, and gratified. We present the mean values of all participants. In Fig. 11, the results show our AR-driven system imposes a significant decrease in workload in cargo handling.

Especially on mental demand, our AR-driven guidance can accurately and intuitively render the virtual cargo on the target location in reality, which drastically saves efforts that users take to keep the cargo location in mind. Moreover, users can work on cargo loading free of nervousness and pressure, which positively affects their work performance and mental health.

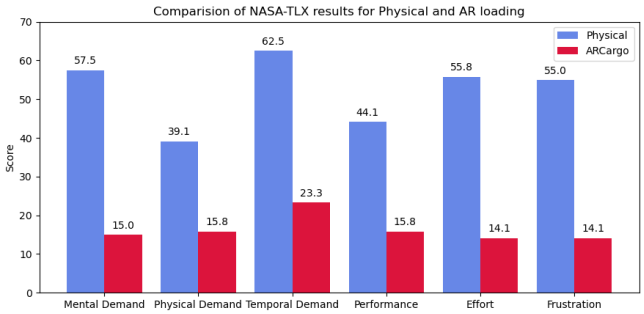

Fig. 11. **Quantitative Results of the NASA-TLX on Physical and AR handling.** The final score of each dimension is the average of scores that participants write on the dimension. Compared to physical handling, the score of six dimensions all decreased significantly on our system, especially on mental demand, which shows our AR-driven system imposes a significant decrease on workload in cargo handling.

V. Conclusions

We present a multi-device integrated cargo loading management system with Augmented Reality (AR), named ARCargo, which can dynamically record the running status of cargo loading in real-time. In the system, we propose a visual localization method using hybrid features to strengthen the localization accuracy. To reduce the burden of storekeepers on cargo loading management, we propose an AR-driven interactive guidance and monitoring alarm program to improve the cargo loading management experience. Extensive experiments show that our system can efficiently, intelligently and conveniently carry out cargo loading management operations.

Acknowledgment

The authors thank Bangbang Yang, Yijin Li, Zhichao Ye and Xinyue Lan for their constructive comments to improve this paper.

References

[1] R. T. Azuma, "A Survey of Augmented Reality," Presence: Teleoperators and Virtual Environments, vol. 6, no. 4, pp. 355–385, 1997.

[2] A. J. Davison, "Real-Time Simultaneous Localisation and Mapping with a Single Camera," in IEEE International Conference on Computer Vision, 2003, pp. 1403–1410.

[3] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," IEEE Transactions on Robotics, vol. 31, no. 5, pp. 1147–1163, 2015.

[4] D. Gálvez-López and J. D. Tardos, "Bags of Binary Words for Fast Place Recognition in Image Sequences," IEEE Transactions on Robotics, vol. 28, no. 5, pp. 1188–1197, 2012.

[5] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-Based Visual–Inertial Odometry Using Nonlinear Optimization," The International Journal of Robotics Research, vol. 34, no. 3, pp. 314–334, 2015.

[6] T. Qin, P. Li, and S. Shen, "VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator," IEEE Transactions on Robotics, vol. 34, no. 4, pp. 1004–1020, 2018.

[7] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in Proceedings of the 7th International Joint Conference on Artificial Intelligence, P. J. Hayes, Ed. William Kaufmann, 1981, pp. 674–679.

[8] A. Irschara, C. Zach, J. Frahm, and H. Bischof, "From Structure-from-Motion Point Clouds to Fast Location Recognition," in IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2599–2606.

[9] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt, "Scalable 6-DoF Localization on Mobile Devices," in European Conference on Computer Vision, 2014, pp. 268–283.

[10] P. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From Coarse to Fine: Robust Hierarchical Localization at Large Scale," in IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12 716–12 725.

[11] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt, "Image Retrieval for Image-Based Localization Revisited," in British Machine Vision Conference, 2012, pp. 1–12.

[12] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating Local Descriptors into a Compact Image Representation," in IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3304–3311.

[13] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 Place Recognition by View Synthesis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 2, pp. 257–271, 2018.

[14] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-NetVLAD: Multi-Scale Fusion of Locally-Global Descriptors for Place Recognition," in IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 14 141–14 152.

[15] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1437–1451, 2018.

[16] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-Supervised Interest Point Detection and Description," in IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 224–236.

[17] P. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning Feature Matching With Graph Neural Networks," in IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 4937–4946.

[18] M. Krogius, A. Haggenmiller, and E. Olson, "Flexible Layouts for Fiducial Tags," in IEEE/RSJ International Conference on Intelligent Robots and Systems, 2019, pp. 1898–1903.

[19] J. Linowes and K. Babilinski, Augmented Reality for Developers: Build Practical Augmented Reality Applications with Unity, ARCore, ARKit, and Vuforia. Packt Publishing, 2017.

[20] A. H. Behzadan, B. W. Timm, and V. R. Kamat, "General-Purpose Modular Hardware and Software Framework for Mobile Outdoor Augmented Reality Applications in Engineering," Advanced engineering informatics, vol. 22, no. 1, pp. 90–105, 2008.

[21] G. Evans, J. Miller, M. I. Pena, A. MacAllister, and E. Winer, "Evaluating the Microsoft HoloLens through an Augmented Reality Assembly Application," in Degraded environments: sensing, processing, and display 2017, vol. 10197, 2017, p. 101970V.

[22] J. L. Schönberger and J.-M. Frahm, "Structure-from-Motion Revisited," in IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4104–4113.

[23] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise View Selection for Unstructured Multi-View Stereo," in European Conference on Computer Vision (ECCV), vol. 9907, 2016, pp. 501–518.

[24] R. I. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, 2nd ed. Cambridge University Press, 2004.

[25] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," CoRR, vol. abs/1704.04861, 2017.

[26] E. Olson, "AprilTag: A Robust and Flexible Visual Fiducial System," in IEEE International Conference on Robotics and Automation. IEEE, 2011, pp. 3400–3407.

[27] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in Advances in psychology, 1988, vol. 52, pp. 139–183.