International Journal of Virtual and Augmented Reality
Volume 6 • Issue 1

Advanced Visual SLAM and Image Segmentation Techniques for Augmented Reality

Yirui Jiang, Cranfield University, UK*
Trung Hieu Tran, Cranfield University, UK
https://orcid.org/0000-0002-3989-4502
Leon Williams, Cranfield University, UK

ABSTRACT

Augmented reality can enhance human perception to experience a virtual-reality intertwined world by computer vision techniques. However, the basic techniques cannot handle complex large-scale scenes, tackle real-time occlusion, and render virtual objects in augmented reality. Therefore, this paper studies potential solutions, such as visual SLAM and image segmentation, that can address these challenges in the augmented reality visualizations. This paper provides a review of advanced visual SLAM and image segmentation techniques for augmented reality. In addition, applications of machine learning techniques for improving augmented reality are presented.

KEYWORDS:

Augmented reality, computer vision, image segmentation, machine learning, visual SLAM

1 INTRODUCTION

Nowadays, augmented reality (AR) has coexisted the real world with virtual objects. The technology has increased human experience in a virtual-reality intertwined world. It has grown in popularity over the last ten years, moving from laboratories into various real-life scenes (Van Krevelen & Poelman, 2010). However, there are numerous issues in AR, leading to the rise of innovations (Masood & Egger, 2019). Outdoor AR systems face challenges such as handling complex large-scale scenes, dealing with real-time occlusion, and rendering virtual objects. Although image segmentation could address the issues, it requires collaboration with other advanced techniques (Roxas et al., 2018). The collaboration is critical in the future trend of AR applications. Since traditional methods for sharing accurate spatial information are insufficient, various machine learning algorithms have been proposed to achieve low-cost and high-efficiency future collaboration systems (Zou et al., 2019). The volume of mobile and industrial AR applications is growing at an exponential rate; however, previous high

DOI: 10.4018/IJVAR.307063
*Corresponding Author

latency, low precision, and unfriendly user experience have hampered the widespread adoption of AR systems (Huang et al., 2013). To overcome these limitations, registration (Hoff et al., 1996), tracking (Runz et al., 2018), image segmentation (Kirillov et al., 2019) and occlusion (Tang et al., 2020) approaches have been proposed with machine learning-improved visualization techniques, i.e., visual SLAM (vSLAM) and image segmentation. This paper provides an overview of the main obstacles of the AR visualizations and their potential solutions. The most recent computer vision technologies are concentrated, particularly machine learning-enhanced innovative technologies.

To date, there exists a lack of comprehensive literature review on the topic of applying the latest machine learning-improved computer vision in AR systems. However, there are some literature reviews related to other research communities in the AR systems, which are shown in Table 1.

As part of the digital transformation of industry, AR improves industrial efficiency, safety, compliance, and costs. Ling et al. (2017) discussed commercial trends in industrial AR. Palmarini et al. (2018) conducted a systematic literature review to identify the most relevant industrial AR technical limitations. Li et al. (2018) examined various Virtual Reality (VR) / AR prototypes, products, and training evaluation paradigms. Gattullo et al. (2020) systematically reviewed the literature on visual assets. Egger et al. (2020) investigated the current challenges and future directions of AR manufacturing. Costa et al. (2022) provided an overview of the current state of the art in AR human-robot collaboration and future development trends. Notably, user research on AR cybersecurity applications is severely lacking. Alzahrani et al. (2022) identified, described, and synthesized research findings on the cybersecurity of the AR industry.

Visual tracking is a fundamental task in AR and has been an active research topic for many years. Kalkofen et al. (2011) pioneered the spatial integration of virtual objects in real-world settings. Rabbi et al. (2012) identified the difficulties of tracking an object in an unfamiliar environment. Billinghurst et al. (2015) investigated general tracking and displaying techniques. Li et al. (2018) reviewed the most recent deep learning-based tracking methods and divided them into three categories based on network structure, functionality, and training. The tracking methods used in AR-based robot maintenance are qualitatively evaluated (Koh et al., 2020). Jiao et al. (2021) reviewed the critical advances made by deep learning, including deep feature representations, network architecture, and four critical issues in visual tracking (e.g., spatiotemporal information integration, target-specific classification, target information update, and bounding box estimation). Zhu et al. (2022) provided an overview of AR visual object tracking on RGB-D videos.

Augmented reality provides users with an engaging, memorable, and impactful interactive experience, resulting in a digital world that closely resembles our physical world and offers a new perspective on reality. Rabbi et al. (2013) concentrated on AR hardware and user experience (UX). Irshad et al. (2014) summarized the UX of AR and identified areas where research was lacking. A thorough and detailed review was presented to assist AR developers in focusing on UX improvement by Irshad et al. (2017). The most influential AR user studies were presented by Dey et al. (2018). Recent AR systems have offered a range of device-specific interaction options and tailored solutions for delivering immersive experiences to users, but with an inherent lack of standardization across devices and applications. To address this issue, a systematic review and evaluation of explicit, task-based interaction methods in immersive environments are presented (Spittle et al., 2022).

The original clunky and poor user experience has been transformed by mobile AR, which you can take with you wherever you go, often on a smartphone device. The number of mobile AR users is increasing as hardware devices improve. AR network implications were highlighted by Westphal et al. (2017). Braud et al. (2017) investigated AR applications and their external infrastructures. Goh et al. (2019) provided context for AR 3D mobile interaction. Lee et al. (2022) reviewed the field of human interaction in connected cities, focusing on AR-driven interaction.

Most existing reviews of AR literature have paid less attention to visualization principles. This paper comprehensively reviews machine learning-enhanced AR visualization applications using vSLAM and image segmentation. The fundamental principles of technology are summarised, and

various novel methods are compared. An extensive literature review is used to investigate existing solutions and potential future works.

The remainder of the paper is organized as follows. Sections 2 and 3 introduce the fundamental principles, novel machine learning-based frameworks, and collaborative architectures of vSLAM.

Table 1. A summary of AR related literature reviews

Reference	Industrial AR					Visual Tracking					User experience				Mobile	
	App.	Par.	Col.	Cha.	Sec.	Gen.	Tra.	Obj.	DL.	Cha.	Gen.	Sta.	Cha.	Imp.	Net.	Int.
Kalkofen et al. (2011)	√					√	√			√						
Rabbi et al. (2012)		√				√		√								
Rabbi et al. (2013)	√			√		√	√			√	√		√	√		
Irshad et al. (2014)											√	√	√	√		√
Billinghurst et al. (2015)	√	√		√		√	√	√		√	√			√		
Braud et al. (2017)	√			√							√		√	√	√	
Irshad et al. (2017)											√	√		√		√
Ling et al. (2017)	√	√		√		√		√		√	√		√	√		√
Westphal et al. (2017)	√	√		√											√	
Dey et al. (2018)	√		√	√		√				√	√	√	√	√		√
Li et al. (2018)	√			√	√								√			
Palmarini et al. (2018)	√			√		√	√			√	√		√	√		√
Goh et al. (2019)	√			√		√				√	√		√	√		√
Egger et al. (2020)	√			√		√				√	√		√			√
Gattullo et al. (2020)	√	√		√		√	√				√					
Koh et al. (2020)	√			√		√	√	√		√						
Jiao et al. (2021)						√			√	√						
Alzahrani et al. (2022)	√			√	√						√		√	√	√	√
Costa et al. (2022)	√		√	√		√	√			√	√		√	√		√
Lee et al. (2022)	√			√		√				√	√		√	√	√	√
Spittle et al. (2022)	√			√		√					√	√	√	√		√
Zhu et al. (2022)	√			√		√	√	√	√	√						

Section 4 provides an overview of image segmentation and its most recent deep learning variants. Section 5 presents using image segmentation to handle environmental understanding and object occlusion in AR applications. Section 6 discusses a fusion AR system that implements advanced visual SLAM and image segmentation to improve current AR performance. Lastly, conclusions are provided in Section 7.Where app. = application, par. = paradigm, col. = collaboration, cha. = challenge, sec. = security, gen. = general, tra. = tracker, obj. = object, dl- = dl-based, std. = standardization, imp. = improvement, net. = network, int. = interaction.

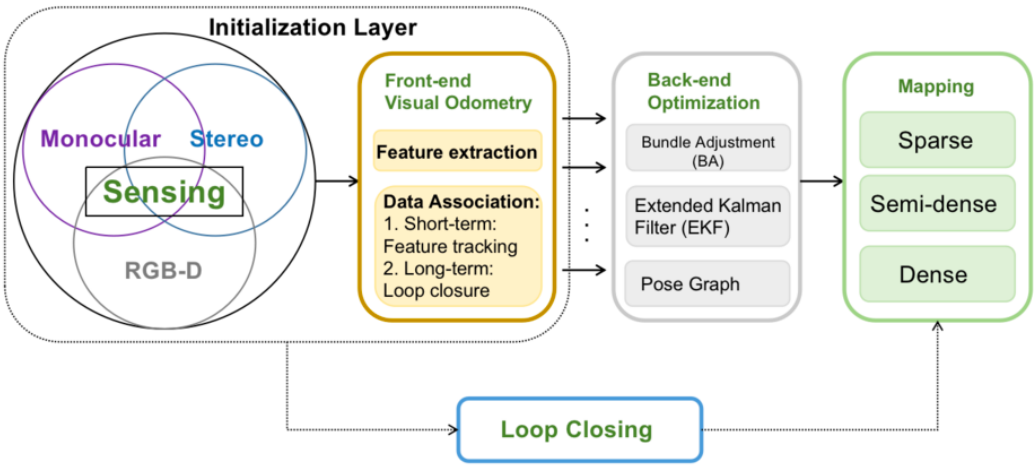
2 A BRIEF OF VISUAL SLAM

This section provides an overview of vSLAM, a technology that uses visual information to calculate the position and orientation of devices and creates maps of an unknown environment with scene structures. The development of machine learning-improved vSLAMs in the past ten years has been classified.

2.1 Basic Framework of vSLAM

The vSLAM framework is composed of five modules: sensing, visual odometry (VO), mapping, optimization, and loop closing (Taketomi et al., 2017) (see Figure1). Sensor data (i.e., monocular, stereo, and RGB-D) are collected by sensing. VO defines a global coordinate system in an unknown environment for estimating camera pose and reconstructing the 3D scene. In VO, feature matching and tracking are for finding the 2D-3D correspondence between 2D images and 3D reconstructed maps, and the perspective-n-Point (PnP) algorithm is for calculating camera pose (Klette et al., 1998; Nistér & Stewénius, 2007). When cameras detect a new environment, 3D structures are expanded and calculated using mapping, and a relocalization module is used to track lost response. The optimization employs the entire map information to reduce accumulative estimate error and make fine-grained adjustments. The current image is cyclically matched with previous ones during loop closing to revise the accumulative error.

Figure 1. Basic framework of vSLAMs



2.2 Various vSLAM Methods

In this section, three main types of vSLAM, such as feature-based (see Table 2), direct (see Table 3) and RGB-D (see Table 4) are discussed (Taketomi et al., 2017).

The most representative feature-based vSLAMs are MonoSLAM (Davison, 2003; Davison et al., 2007), PTAM (Klein & Murray, 2007) and ORB-SLAM (Mur-Artal & Tardós, 2017). MonoSLAM is the first monocular vSLAM (Davison, 2003; Davison et al., 2007), using an extended Kalman filter (EKF) to estimate the feature points and camera motions. However, MonoSLAM is rarely used in real-time applications since its computational cost is increased in an unknown environment. Therefore, the PTAM is proposed (Klein & Murray, 2007) with using bundle adjustment (BA) for optimization. It separates tracking and mapping to different CPUs and executes them in parallel, and runs the feature point computing on different threads (Nistér, 2004; Strasdat et al., 2012; Williams et al., 2007), making the tracking of camera motion in real-time possible. Several multi-threading approaches have been proposed in light of the PTAM's good performance. Castle et al. (2008) developed a version of multiple maps of the PTAM. Klein et al. (2009) enhanced the PTAM for mobile phones. To solve the local minimum dilemma in the PTAM (Mei et al., 2009), Mur-Artal et al. (2014) proposed ORB-SLAM. ORB-SLAM has become the most widely used feature-based monocular vSLAM in the literature and industrial products by expanding stereo vSLAM with RGB-D vSLAM of sensing and combining BA with 7 DoF pose graph in optimization (Mur-Artal & Tardós, 2017). The performance of current visual SLAM algorithms degrades significantly in dynamic scenarios due to the disturbance of dynamic objects. Cheng et al. (2019) proposed a novel method that uses optical flow to distinguish and eliminate dynamic feature points from extracted ones when RGB images are used as the only input to address this issue. Liu et al. (2019) presented an energy-efficient architecture for a real-time ORB-based visual SLAM system by accelerating the most time-consuming stages of feature extraction and matching on the FPGA platform. Kiss-Illés et al. (2019) proposed GPS-SLAM, an augmented version of Oriented FAST and Rotated BRIEF feature detector ORB-SLAM that uses GPS and inertial data to make the algorithm capable of dealing with low frame rate datasets. Mur-Artal & Tardós (2017) developed a benchmark method in the ORB-SLAM domain. However, it consumes significant time for computing descriptors that never get reused unless a frame is selected as a keyframe. To overcome these problems, Fu et al. (2022) presented FastORB-SLAM, which is lightweight and efficient as it tracks key points between adjacent frames without computing descriptors.

Feature-based vSLAMs perform well in simple scenes, but fall short in complex environments (Eade & Drummond, 2006; Hirose & Saito, 2012). Direct vSLAMs have been proposed for a variety of scenarios, with photometric consistency used to measure error rather than feature points or descriptors. The first direct vSLAM, DTAM (Newcombe et al., 2011), is popular in mobile applications (Ondrúška et al., 2015). It registers input images with reconstructed 3D maps, and uses depth information to optimize spatial continuity (Okutomi & Kanade, 1993; Rudin et al., 1992). Stühmer et al. (2010) extended the DTAM by analyzing the depth information for each pixel, and tracking with the PTAM (Klein & Murray, 2007). The LSD-SLAM (Engel et al., 2014; Caruso et al., 2015; Engel et al., 2015), a first semi-dense vSLAM system, was proposed to explore textureless areas. Following the success of semi-dense VO (Engel et al., 2013), the LSD-SLAM randomly initializes the depth information of each pixel and only reconstructs the high-intensity gradient areas (Schöps et al., 2014). Forster et al. (2014) also proposed semi-direct VO (SVO) as a sparse version of DTAM and LSD-SLAM by combining feature-based and direct methods. Engel et al. (2017) pioneered the direct sparse odometry (DSO). This is a fully direct method that divides images into several blocks as inputs rather than uses whole images as SVO. Triputen et al. (2018) proposed methods to improve the performance metrics of a visual SLAM system. They analyzed the accuracy of 3D environments reconstructed using an LSD-SLAM system that mounts a monocular camera on a fixture of a collaborative robot. Outahar et al. (2021) proposed a solution to combine direct and indirect methods in order to increase robustness and allow AR to move seamlessly between different types of scenes.

To capture efficient and convenient information, RGB-D cameras (Geng, 2011) (i.e., Microsoft Kinect (Zhang, 2012)) are used for RGB-D vSLAM. It estimates camera motion and constructs 3D reconstruction using depth information and iterative closest point (ICP) (Besl, 1992). There are several voxel-related works in RGB-D vSLAM. KinectFution (Newcombe et al., 2011) used depth maps in voxel space to reconstruct the 3D structure, whereas Kähler et al. (2015) used hashing mapping to reduce computational cost, leading to generalizability in mobile applications. An object-level RGB-D vSLAM is proposed (Salas-Moreno et al., 2013) by reconstructing accurate maps through pre-registered 3D objects without large amounts of data. Tateno et al. (2016) demonstrated a real-time algorithm for segmenting instances and labelling objects. With the growing popularity of RGB-D cameras, an increasing number of AR devices such as Google Tango2 and Structure Sensor3 provide RGB-D vSLAM APIs with stable estimation results, which will promote the innovation of RGB-D vSLAM. Yao et al. (2018) proposed a novel VO for RGB-D cameras based on edges and points, which they implemented on the TUM RGB-D dataset to achieve more accurate and stable localization in dynamic environments. Alves et al. (2020) proposed a wireless remote RGB-D visual SLAM solution for robots with low computational power. Dai et al. (2021) developed a method for removing the influence of moving objects in dynamic environments.

2.3 Improving vSLAM with Machine Learning

Traditional features-based vSLAM algorithms are sensitive to changing lighting, dynamic targets, and under-textured environments. There are longstanding problems with scale-invariant feature transform (SIFT), or oriented FAST and rotated BRIEF (ORB) based approaches. Convolutional neural networks (CNNs) have been successful at learning optimal image feature representation. To

Table 2. Feature-based vSLAMs

Reference	System	Map density	Global optimization	Loop closure	Contributions
Davison et al. (2003)	MonoSLAM	Sparse			The first monocular vSLAM
Klein et al. (2007)	PTAM	Sparse	√		The first vSLAM with BA
Castle et al. (2008)	Extended PTAM	Sparse	√		A multiple maps version of PTAM
Klein et al. (2009)	Multi-Maps PTAM	Sparse	√		Extended PTAM for mobile devices
Stühmer et al. (2010)	Extended DTAM	Sparse			Analyze pixels depth information
Mur-Artal et al. (2014)	ORB-SLAM	Sparse	√	√	Widely used monocular vSLAM
Mur-Artal et al. (2017)	ORB-SLAM2	Sparse	√	√	Expand ORB-SLAM with RGB-D
Cheng et al. (2019)	Extended ORB-SLAM	Sparse	√	√	Improve ORB-SLAM for dynamic environment
Liu et al. (2019)	ESLAM	Sparse	√	√	Energy-efficient real-time ORB-SLAM
Kiss-Illés et al. (2019)	GPS-SLAM	Sparse	√	√	Augmented-enhanced ORB-SLAM
Fu et al. (2022)	FastORB-SLAM	Sparse	√	√	Lightweight and efficient ORB-SLAM

Table 3. Direct vSLAMs

Reference	System	Map density	Global optimization	Loop closure	Contributions
Newcombe et al. (2011)	DTAM	Dense			The first direct vSLAM
Engel et al. (2014)	LSD-SLAM	Semi-dense	√	√	The first semi-dense vSLAM
Forster et al. (2014)	SVO	Sparse			Extend DTAM and LSD-SLAM
Engel et al. (2017)	DSO	Sparse			A fully direct method
Triputen et al. (2018)	Extended LSD-SLAM	Sparse	√	√	Collaborative robot LSD-SLAM for scaled 3D environment
Outahar et al. (2021)	Fusion LSD-SLAM & ORB-SLAM2	Sparse	√	√	Extend LSD-SLAM & ORB-SLAM2

Table 4. RGB-D vSLAMs

Reference	System	Map density	Global optimization	Loop closure	Contributions
Newcombe et al. (2011)	KinectFusion	Dense			Use depth maps in voxel space
Salas-Moreno et al. (2013)	SLAM++	Dense	√	√	An object level RGB-D vSLAM
Kähler et al. (2015)	Hashing KinectFusion	Dense			Extend hashing KinectFusion
Tateno et al. (2016)	Extend 2.5D vSLAM	Dense	√		A real-time 2.5D semantic system
Yao et al. (2018)	Dynamic RGB-D vSLAM	Dense	√	√	A RGB-D vSLAM for dynamic environments
Alves et al. (2020)	Remote RGB-D vSLAM	Dense	√	√	A wireless remote solution for low computational powered robots
Dai et. al. (2021)	Dynamic RGB-D vSLAM	Dense	√	√	A solution for eliminating moving objects influence

avoid feature extraction and matching, some recent solutions have applied deep learning (DL) in VO and loop closure. Konda et al. (2015) developed an end-to-end deep neural network (DNN) for camera speed and direction prediction. Costante et al. (2015) used CNNs to deal with image motion blur and lighting changes. Handa et al. (2016) extended spatial transform network (Jaderberg et al., 2015) for end-to-end VO and image depth estimation. For monocular VO, an end-to-end, sequence-to-sequence probabilistic visual odometry (ESP-VO) framework based on deep recurrent convolutional neural networks (RCNN) is proposed (Wang et al., 2018). Almalioglu et al. (2018) presented a generative unsupervised learning framework. It predicts 6-DoF pose camera motion and monocular depth map of the scene from unlabelled RGB image sequences using deep convolutional Generative Adversarial Networks (GANs). Wang et al. (2020) reviewed approaches, challenges, and applications for deep

visual odometry to understand better how DL can profit and optimize the VO systems. Traditional VO systems are unable to operate well in challenging environments. To address this issue, Zhang et al. (2021) combined the classical stereo VO system's multi-view geometry constraints with the robustness of DL to present an unsupervised pose correction network for the classical stereo VO system. Single sensor approaches are frequently prone to failure due to degraded image quality caused by environmental factors, such as camera placement. Kaygusuz et al. (2021) proposed a deep sensor fusion framework that estimates object motion using both pose and uncertainty estimations from multiple onboard cameras to address this issue. Qin et al. (2021) proposed a camera-based localization method for tracking and recording the scanner position in real-time and providing a DL-based segmentation method.

Most solutions combine neural networks in loop closing modules to fully exploit the high recognition rate of DL. Chen et al. (2014) implemented a novel pre-trained CNN network for extracting image feature. To tackle the sensitivity problem of complex scenes, Hou et al. (2015) extended AlexNet (Krizhevsky et al., 2012) to improve system robustness of illumination changes. Gao et al. (2015) provided an auto-encoder to extract image features for image matching. Arandjelovic et al. (2016) proposed an end-to-end scene recognition algorithm based on vector of locally aggregated descriptors (VLAD). Qiu et al. (2018) demonstrated the Siamese-ResNet network, which combines the Siamese network with the ResNet network to detect loop closure. Duan et al. (2019) investigated the most recent research advances in DL-based vSLAM loop closure. Liu et al. (2019) used the DL point cloud description and the coarse-to-fine sequence matching strategy to solve the loop-closure detection problem. While these approaches are successful in many applications, they do not use all the information that a monocular image provides, and many of them, in particular, require user-chosen thresholding to close loops, which may impact generality in practical applications. Merrill et al. (2019) addressed these concerns by extracting all three modes of information from a custom deep CNN trained specifically for place recognition. Memon et al. (2020) proposed a novel approach based on a super dictionary that is distinct from the traditional BoW dictionary and employs more advanced and abstract DL features. Chen et al. (2021) presented a novel end-to-end loop-closure detection method based on continuous learning. Continuous learning can effectively suppress the decline of the memory capability of a simultaneous localization and mapping system. The proposed system incorporates the orthogonal projection operator into the loop-closure detection to overcome the catastrophic forgetting problem of mobile robots in large-scale and multi-scene environments. Arshad et al. (2021) conducted a comprehensive review of the existing literature on loop closure detection algorithms for visual and Lidar SLAM and discussed their insights and limitations.

With the innovative technologies based on machine learning, vSLAM is more robust to environmental changes (i.e., light and seasons). However, there are still issues with hidden layer selection, neural network architectures, and network parameter optimizations. Table 5 summarizes the most recent vSLAM-DL works. Table 6 shows a comparison of traditional vSLAMs and DL-improved vSLAMs.

3 COLLABORATIVE VSLAM FOR AR SYSTEMS

With the growing popularity of mobile AR devices (Shafi et al., 2017), multi-agent vSLAM systems, also known as collaborative vSLAMs, have received a lot of attention in recent years. However, only a few of vSLAMs have been successfully applied with multiple agents. The basic architecture of collaborative vSLAMs is introduced in this section, as is the information sharing process between different agents.

3.1 Basic Architecture of Collaborative vSLAM for AR Systems

Collaborative AR systems display virtual objects on multiple devices in real time, and unify the users' coordinate systems by sharing different user statuses. Although collaborative vSLAMs use modules

Table 5. A summary of deep learning in vSLAM

Reference	Module		Contributions
	VO	Loop closing	
Chen et al. (2014)		√	CNN for feature extraction
Gao & Zhang et al. (2015)		√	Introduce auto-encoder
Hou et al. (2015)		√	Tackle illumination changes
Konda et al. (2015)	√		DNN to predict camera pose
Costante et al. (2015)	√		Focus on complex scenarios
Arandjelovic et al. (2016)		√	VLAD based scene recognition
Handa et al. (2016)	√		Extended spatial network
Almalioglu et al. (2018)	√		Generative unsupervised learning framework
Qiu et al. (2018)		√	A novel Siamese-ResNet network
Wang et al. (2018)	√		Directly infer pose and uncertainty with RCNN
Duan et al. (2019)		√	Review most recent research
Liu et al. (2019)		√	Solve detection problem
Memon et al. (2020)		√	A super dictionary-based solution
Merrill et al. (2019)		√	Address user-chosen practical issues
Wang et. al. (2020)	√		Review approaches, challenges, and applications for deep VO
Arshad et al. (2021)		√	A comprehensive review of insights and limitations
Chen et al. (2021)		√	End-to-end loop-closure detection
Kaygusuz et al. (2021)	√		Deep sensor fusion VO framework
Qin et al. (2021)	√		DL-based segmentation
Zhang et al. (2021)	√		Unsupervised pose correction network

Table 6. Comparison of traditional vSLAM and deep learning-based vSLAM

Task	Traditional	Deep learning
Parameters	Small data scale,	Large data scale,
	Short tuning cycle	Long tuning cycle
Intelligibility	Good interpretability	Poor interpretability
Generalization	Insufficient information,	Sufficient information,
	Few parameters,	Many parameters,
	Weak generalization	Strong generalization
Adaptability	Weak transfer capacity	Strong transfer capacity
Design flow	Separate design and training	Design and training simultaneously

similar to traditional vSLAMs, they are more focused on collaborative pose estimation, collaborative mapping, place recognition, and relative transform computation. Some collaborative AR systems use a centralized architecture with a user front end and a server back end. The front-end computes

real-time user data, and the back end perform map fusion, simplification, and optimization. However, collaborative AR systems are gradually shifting to more decentralized architectures (Schmuck & Chli, 2019; Karrer et al., 2018), outsourcing computational tasks to each agent, which greatly reduces the computational cost of the central server.

3.2 Different Methods of Collaborative vSLAM for AR Systems

The mapping module distinguishes collaborative vSLAM systems from traditional vSLAM systems. This section discusses one global map and multiple local maps-based collaborative vSLAMs.

In the one global map method, two cameras are placed as close as possible to unify the coordinates of different AR devices (Klein & Murray, 2007). Examples include PTAMM (Castle et al., 2008), CoSLAM (Zou & Tan, 2012), Multi-robot CoSLAM (Perron et al., 2015), and so on. PTAMM systems use the PTAM algorithm to collect data from two different camera trackers and transmitting it to the same map (Castle et al., 2008). To apply in larger-scale scenes, CoSLAM (Zou & Tan, 2012) systems divide twelve independent cameras into multiple groups. The multi-robot CoSLAM systems (Perron et al., 2015) control moving targets based on real-time camera poses and 3D feature trajectories, laying the groundwork for future collaborative AR systems. Zhang et al. (2018) proposed a distributed and collaborative monocular simultaneous localization and mapping system for multi-robot systems operating in large-scale environments where monocular vision is the only exteroceptive sensor. A homogeneous mobile robot team will map unknown indoor environments collaboratively. Hentout et al. (2020) proposed a distributed multi-agent coordination approach for mapping to provide a comprehensive view of the entire environment. Jang et al. (2021) proposed a collaborative monocular SLAM with a map fusion algorithm that takes advantage of rendezvous, which can occur when multi-robot team members operate nearby. While traditional approaches use powerful cloud servers to accelerate SLAM computing, the significant communication overhead prevents real-world implementation. Huang et al. (2021) presented ColaSLAM, a multi-robot collaborative laser SLAM system with robot-edge synergy, to address this challenge. SwarmMap, a framework design that scales up collaborative vSLAM service in edge offloading settings, was demonstrated and implemented by Xu et al. (2022). In general, collaborative vSLAMs on one global map transmit dispersed data for centralized and unified preprocessing. However, as the number of cameras increases, the communication load of one global map collaborative vSLAMs becomes prohibitively high, which leads to challenges.

Since one global map collaborative vSLAMs wastes the computation power of other agents, multiple local maps collaborative vSLAMs are proposed. Multiple local maps collaborative vSLAMs compute data in distributed agents and only transmit the intermediate result to the server, which greatly reduces the communication load. In C2TAM (Riazuelo et al., 2014) system, each agent separately sends keyframe images instead of all video data to the server for fusion and 3D reconstruction. However, since intermediate data in C2TAM (Riazuelo et al., 2014) is still too large, CSfM (Forster et al., 2013) models were proposed. CSfM (Forster et al., 2013) models only send the feature points and relative poses of keyframe images to the server. Furthermore, CCM-SLAM systems (Schmuck & Chli, 2019) only use a portion of keyframes rather than all of them. Although these works are real-time, they are ineffective in complex environments. To improve robustness and accuracy, Keivan et al. (2016) proposed the MOARSLAM system, which combines visual sensors with monocular visual-inertial SLAM. CVI-SLAM (Karrer et al., 2018) extended MOARSLAM (Keivan et al., 2016) to mobile devices, providing foundation for collaborative AR systems. Opdenbosch et al. (2019) investigated data efficiency for visual information exchange in a collaborative visual SLAM setup. Liu et al. (2021) proposed a centralized multi-intelligence collaborative monocular visual-inertial SLAM deployed on multiple iOS mobile devices. COVINS (Schmuck et al., 2021) is a novel collaborative SLAM system that supports multi-agent, scalable SLAM in large environments and for large teams of more than ten agents. Although sharing gestures and gazes can improve AR remote collaboration, most current systems only allow collaborators to share image data. Wang et al. (2022) describe a novel remote

collaborative platform based on 2.5D gestures and gaze to address this issue (2.5DGG). In general, multiple local maps-based collaborative vSLAMs process data independently, and only transmit intermediate data to the server, which reduces the communication load significantly. The architectures of centralized and decentralized multi-agent vSLAMs are summarized in Table 7.

4 A BRIEF OF IMAGE SEGMENTATION WITH DEEP LEARNING

Image segmentation is widely used in autonomous driving (Ess et al., 2009; Geiger et al., 2012; Cordts et al., 2016), human-machine interaction (Oberweger et al., 2015) and augmented reality as

Table 7. A list of collaboration vSLAMs

Reference	System	Type		Description
		One global map	Multiple local maps	
Castle et al. (2008)	PTAMM	√		Transmit data to a same map with two camera trackers
Zou et al. (2012)	CoSLAM	√		Apply it to larger scale scenes with twelve cameras
Forster et al. (2013)	CSfM		√	Send feature points on key frame images to the server
Riazuelo et al. (2014)	C2TAM		√	Send key frame images to the server
Perron et al. (2015)	Multi-robot CoSLAM	√		Control moving targets as a real time system
Keivan et al. (2016)	MOARSLAM		√	Combines visual sensor with monocular visual-inertial SLAM
Karrer et al. (2018)	CVI-SLAM		√	Extend MOARSLAM to mobile devices
Zhang et al. (2018)	Multi-robot CoSLAM	√		Apply for large-scale environments
Opdenbosch et al. (2019)	Large-scale CoSLAM		√	Multiple agents for unknown large-scale environments
Schmuck et al. (2019)	CCM-SLAM		√	Only uses part of key frames instead of all
Hentout et al. (2020)	Homogeneous CoSLAM	√		A homogeneous multi-mobile solution
Huang et al. (2021)	ColaSLAM	√		A robot-edge synergy solution
Jang et al. (2021)	Monocular CoSLAM	√		A rendezvous multi-robot solution
Liu et al. (2021)	Mobile CoSLAM		√	Multiple mobile solution
Schmuck et al. (2021)	COVINS		√	A large team solution
Wang et al. (2022)	2.5DGG CoSLAM		√	Share gestures and gaze for remote collaboration
Xu et al. (2022)	SwarmMap	√		ColaSLAM in edge offloading settings

a solution to environmental understanding. DL-based image segmentation is a process from coarse to fine, represented by AlexNet (Krizhevsky et al., 2012), VGG-16 (Simonyan & Zisserman, 2014), GoogLeNet (Szegedy et al., 2015), and ResNet (He et al., 2016).

The Fully Convolutional Network (FCN) was proposed to learn hierarchical structure and feature point information (Long et al., 2015). FCN produces spatial maps rather than classification scores (Zeiler et al., 2011; Zeiler & Fergus, 2014), which improves segmentation accuracy and standard data set efficiency. FCN, on the other hand, under-utilizes global context information, performs poorly in real-time at high resolution, and is unsuitable for unstructured data. To obtain a real-time semantic system, SegNet (Badrinarayanan et al., 2017) used an encoder and decoder, followed by a softmax classifier, to pixel-wise label output. Simultaneous Detection and Segmentation (SDS) (Hariharan et al., 2014) extended FCN as a bottom-up hierarchical image segmentation system. Furthermore, Pinheiro et al. (2015) presented the DeepMask model, which is based on a single ConvNet. Zagoruyko et al. (2016) provided Fast R-CNN on MultiPath classifier. Uhrig et al. (2019) proposed Box2pix with FCN to predict object bounding boxes, as well as efficient and straightforward post-processing to assign each object pixel to its best match for object detection. Yolact++ was proposed by Boliah et al. (2020) for real-time instance segmentation, with two parallel subtasks that linearly combine prototypes with mask coefficients to generate instance masks. This method can produce high-quality masks with time stability because it does not rely on re-pooling. SOLOv2 (Wang et al., 2020) is a straightforward, simple, and quick framework for segmentation. It dynamically segments each instance in an image without relying on bounding box detection, making it efficient and comprehensive. SOLOv2 is very useful for panoramic segmentation. Tian et al. (2021) proposed a high-performance method for implementing mask-level instance segmentation with only bounding box annotations. Instead of the traditional crude manual initialization, E2EC (Zhang et al., 2022) is a learnable contour-based instance segmentation method (see Table 8).

To achieve temporal continuity between video data, Shelhamer et al. (2016) presented a clockwork FCN model based on the feature velocity with different update rates to handle layers of varying depths. Zhang et al. (2014) learned hierarchical features from multi-channel inputs using 3D CNN. Tran et al. (2016) developed a deep end-to-end voxel-to-voxel prediction system. Li et al. (2018) created a video semantic segmentation framework that includes a feature propagation module to adaptively fuses features over time via spatial variant convolution, lowering the cost per frame computation. Paul et al. (2020) proposed an efficient video segmentation (EVS) pipeline that combines a high-speed optical flow method for exploiting temporal aspects of video and propagating semantic information from frame to frame. Hu et al. (2020) presented TDNet, a temporally distributed network designed for video semantic segmentation that is fast and accurate. Wang et al. (2021) proposed a temporal memory attention network (TMANet) based on a self-attentive mechanism. It integrates long-range temporal relationships over video sequences without requiring exhaustive optical flow prediction. Alapatt et al. (2021) developed a temporally constrained neural network (TCNN) as a semi-supervised framework for surgical video semantic segmentation. Standard images or videos typically contain only a small number of semantic categories throughout the label set. He et al. (2022) proposed decomposing segmentation into two subproblems, image-level or video-level multi-label classification and pixel-level particular label classification (see Table 9).

The FCN methods have achieved a good performance on 2D image segmentation and 3D voxel handling. Huang et al. (2016) addressed the 3D point cloud labeling challenge using 3D CNNs, which reduce a priori labeling knowledge without using hand-crafted features. Qi et al. (2018) proposed a framework for learning directly in the raw point cloud that allows for accurate estimation of 3D bounding boxes even in the presence of solid occlusions and sparse points. 3D-SIS (Hou et al., 2019), a novel 3D semantic instance segmentation in RGB-D scanning that learns jointly from geometric and color signals, predicts object bounding boxes, class labels, and instance masks with high accuracy. Wu et al. (2019) proposed a point cloud segmentation method based on deep learning with improved model structure, training losses, and additional input channels. Manually annotating complex point

cloud datasets is both time-consuming and error-prone. SnapshotNet (Li et al., 2022) was proposed as a self-supervised feature learning method for rapidly learning practical features from unlabeled large-scale point cloud data (see Table 10).

The local and global information is balanced by image segmentation-based AR applications. Refinement CNNs, such as Conditional Random Fields (CRFs), dilated convolutions, and multi-scale aggregation, use the global context of images to solve local ambiguities. CRFs improve local information and enable the capture of long-range dependencies (Rother et al., 2004; Shotton et al., 2009). The DeepLab models (Chen et al., 2014; Chen et al., 2017) extend previous CRF model (Krähenbühl & Koltun, 2011; Krähenbühl & Koltun, 2013), considering both short and long-range interactions. The CRFasRNN system (Zheng et al., 2015), fully training an end-to-end system with a combination of the CRF and the FCN. Kalogerakis et al. (2017) presented a deep architecture for segmenting 3D objects into labeled semantic parts, combining FCN and CRF to produce coherent 3D shape segmentation and achieving promising segmentation results on noisy 3D shapes. Zhang et al. (2018) proposed a framework for semantic scene segmentation and completion optimization using dense CRF based on a single depth image. Ji et al. (2020) demonstrated a cascaded CFR framework inspired by the skip connection of FCN to learn boundary information from multiple layers, thereby improving model boundary localization and supplementing image semantic information (see Table 11).

To expand receptive fields without losing resolution, dilated convolutions are proposed. Yu et al. (2015) proposed a multi-scale context aggregation module, and Paszke et al. (2016) developed a real-time network ENet that increase dilation rates without additional cost. Multi-scale aggregation was proposed to make use of multiple networks with different scales. Raj et al. (2015) introduced a VGG-16 based multi-scale system that processes a shallow convolutional network on one path and a fully convolutional VGG-16 on the other. Roy et al. (2016) extracted features from coarse-to-fine on four multi-scale CNNs. Bian et al. (2016) trained each network independently in an FCNs architecture with fine-tuning the last layer. MFR-CNN (Zhang et al., 2018) is a novel CNN model that represents local and global features at multiple scales. It can detect real-world traffic objects accurately and performs well with heavily occluded objects. Lin et al. (2018) proposed a novel multi-scale context intertwining (MSCI) framework for aggregating features at different scales to achieve accurate multi-scale semantic segmentation. Through the connection of two LSTM chains, the traditional one-way propagation of information is combined with feature graph pairs in a bidirectional and circular manner, improving the effectiveness of feature extraction. FgSegNet (Lim et al., 2020) extracts multi-scale features from images and generates a robust feature pool to segment moving objects from the background accurately. Computational resources limit real-time semantic segmentation, but performing multi-scale contextual aggregation can solve this problem within a limited computational budget. Farsee-net (Zhang et al., 2020) includes a cascaded factorized atrous spatial pyramid pooling (CF-ASPP) module, a lightweight cascade structure that improves runtime efficiency through contextual information. Gao et al. (2021) proposed a new lightweight network using a multi-scale context fusion (MSCFNet) scheme to enhance feature representation and improve segmentation efficiency to strike a good balance between semantic segmentation accuracy, inference speed, and model size. Hui et al. (2022) proposed a novel adaptive segmentation model for multi-scale targets based on the DeepLabv3+ framework to improve the segmentation accuracy of small and obstructed image targets. The model reduces the number of model parameters and model size while improving segmentation speed and performance. In the integrating context knowledge method, researchers have expanded the traditional CNNs architecture with innovation and considered both local and global information (see Table 12).

5 AR SYSTEMS BASED ON IMAGE SEGMENTATION

This section discusses image segmentation-based AR systems and their future works. To handle real-time occlusion in AR systems, the foreground masks of real RGB scene (Rother et al., 2004; Li et al., 2004; Hasinoff et al., 2006; Criminisi et al., 2006; Kakuta et al., 2008) and RGBD images (Zhu et al.,

Table 8. FCN-based image segmentation

Reference	Name	Category			Architecture	Contributions
		General	Decoder variants	Instance Segmentation		
Hariharan et al. (2014)	SDS			√	R-CNN + Box CNN	A bottom-up image segmentation system
Kendall et al. (2015)	Bayesan SegNet		√		VGG-16 + Decoder	Uncertainty modeling
Long et al. (2015)	the FCN	√			VGG-16	Use spatial maps instead of scores
Pinheiro et al. (2015)	Deep Mask			√	VGG-A	Extend single ConvNet
Zagoruyko et al. (2016)	MultiPathNet			√	Fast R-CNN + Deep Mask	MultiPath classifier
Badrinarayanan et al. (2017)	SegNet		√		VGG-16 + Decoder	Encoder-decoder
Uhrig et al. (2019)	Box2pix			√	GoogLeNet	Single-shot pixel to box
Bolya et al. (2020)	Yolact++			√	RetinaNet	Fully-convolutional real-time model
Tian et al. (2021)	BoxInst			√	Mask R-CNN	Projection and pair-wise affinity mask loss
Wang et al. (2020)	SOLOv2			√	Mask R-CNN	Fast dynamic solution
Zhang et al. (2022)	E2EC			√	DML	A novel contour-based method

Table 9. Video sequences-based image segmentation

Reference	Name	Architecture	Contributions
Zhang et al. (2014)	3DCNN-Zhang	3DCNN	Learn features from multi-channel
Shelhamer et al. (2016)	Clockwork Convent	FCN	A clockwork FCN model
Tran et al. (2016)	End2EndVox2Vox	C3D	A deep end-to-end voxel-to-voxel system
Li et al. (2018)	Low-Latency VS	ResNet-101	Design a video segmentation framework
Paul et al. (2020)	EVS	ICNet	High-speed optical flow method
Hu et al. (2020)	TDNet	APM	A temporally distributed network
Wang et al. (2021)	TMANet	TMA	Self-attention mechanism
Alapatt et al. (2021)	TCNN	TCNN	A semi-supervised framework for surgical
He et al. (2022)	MLSeg	Query2Label	Generic two sub-problems framework

2008; Kanbara et al., 1999; Kim et al., 2003; Kim et al., 2008) are created. RGBD images are used to estimate the foreground accurately, while RGB images are used to create depth maps (Hebborn et al., 2017; Du et al., 2016; Fukiage et al., 2014). For example, the grassland is identified as background by

Table 10. 3D data-based image segmentation

Reference	Name	Architecture	Contributions
Huang et al. (2016)	Huang-3DCNN	3DCNN	Voxelized point clouds
Qi et al. (2018)	Point Net	T-Net	Use unordered point sets
Hou et al. (2019)	3D-SIS	3D-RPN	RGB-D scan for 2D and 3D feature learning
Wu et al. (2019)	Squeezesegv2	SqueezeSeg	3D point cloud segmentation
Li et al. (2022)	SnapshotNet	SVM	Work on unlabeled complex 3D scene

Table 11. CRFs-based image segmentation

Reference	Name	Architecture	Contributions
Chen et al. (2014)	DeepLab	VGG-16/ ResNet-101	Consider short and long-time interactions
Zheng et al. (2015)	CRFasRNN	FCN-8s	Combine the CRF and the FCN
Kalogerakis et al. (2017)	3DCRF	FCN	Combine the CRF and the FCN for 3D
Zhang et al. (2018)	Dense CRF	SSCNet	Dense CRF based on single depth image
Ji et al. (2020)	Cascaded CRF	FCN-8s	Cascaded CRF method with FCN

semantic segmentation, therefore, is placed behind other instances. For tackle real-time AR occlusion challenges, Zhao et al. (2018) and Zhou et al. (2017) combined foreground probability map with semantic segmentation for blending and adopted real-time CNN-based image segmentation for a better environmental understanding in AR systems. Image segmentation that affects significantly indoor and outdoor AR applications (Roxas et al., 2018) has two limitations: time consumption and generalisation. Per-frame semantic segmentation is time-consuming and makes insufficient use of input data continuity and temporal information. For generalization, the classification range of the existing segmentation system is narrow and coarse. Future work could include fine-tuning blending parameters to handle complex scenes (Roxas et al., 2018) and investigate more specific AR scenarios.

6 AR SYSTEMS BASED ON VSLAM AND IMAGE SEGMENTATION

Collaborative vSLAM achieves good performance in static environments (Dou et al., 2016; Innmann et al., 2016; Newcombe et al., 2015; Zollhöfer et al., 2014; Rünz & Agapito, 2017), but lacks environmental understanding with purely geometric output and is prone to characteristic errors in dynamic scenes (Castle et al., 2007; Civera et al., 2011; McCormac et al., 2017; Salas-Moreno et al., 2013; Tateno et al., 2017). To take full advantage of fusing object labels into the map, combining vSLAM and image segmentation leads to real-time tracking and detailed reconstruction (Runz et al., 2018). These fusion AR systems integrate real-time vSLAM architecture with image segmentation to reconstruct the environment while also understanding the scene and sharing information within a group (see Figure 2). Fusion AR systems reconstruct multiple objects with accurate semantic labels, predict object labels with high boundary accuracy (He et al., 2017), track multiple independently moving objects at the same time, improve system accuracy with real-time semantic information (Kaiming et al., 2017; Rünz & Agapito, 2017), and use semantics for tracking reconstruction (Rünz & Agapito, 2017). Some fusion AR systems track rigid objects using a dynamic semantic-based method (Runz et al., 2018). In the field of autonomous driving, DynSLAM (Kaiming et al., 2017)

Table 12. Dilated convolutions & multi-scale prediction-based image segmentation

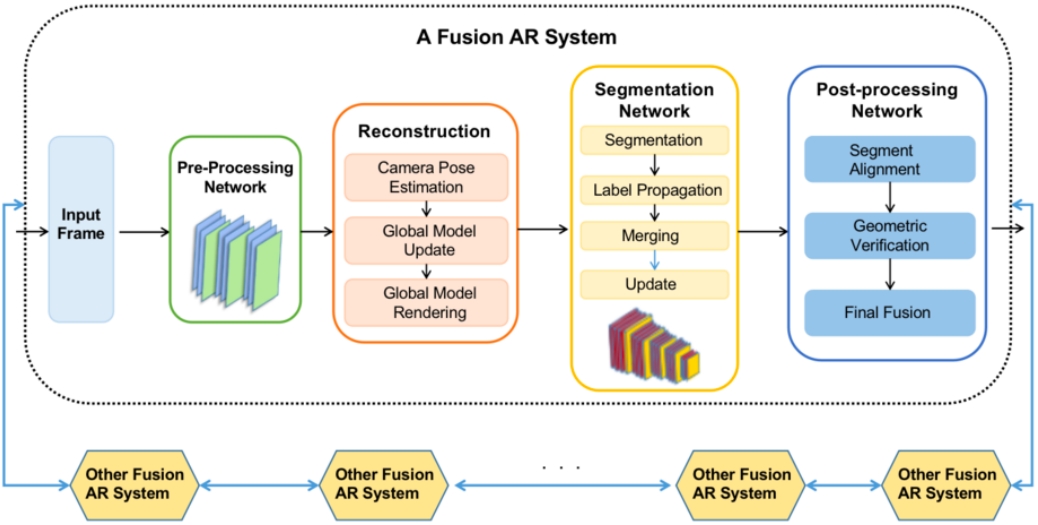
Reference	Name	Category		Architecture	Contributions
		Dilated Convolutions	Multi-scale Prediction		
Raj et al. (2015)	Multi-scale-CNN-Raj		√	Multi-scale-CNN	Introduced a multi-scale system
Yu et al. (2015)	Dilation	√		VGG-16	Multi-scale context aggregation
Bian et al. (2016)	Multi-scale-CNN-Bian		√	Multi-scale-CNN	Independently multi-scale FCNs
Paszke et al. (2016)	zENet	√		ENet bottleneck	Increase dilation rates
Roy et al. (2016)	Multi-scale-CNN-Roy		√	Multi-scale-CNN	A four multi-scale CNNs
Bian et al. (2016)	Multi-scale-CNN-Bian		√	Multi-scale-CNN	Independently trained multi-scale FCNs
Lin et al. (2018)	Multi-scale-CNN-Lin		√	Multi-scale-CNN	Multi-scale context intertwining framework
Zhang et al. (2018)	MFR-CNN		√	VGG/ResNet	Multi-scale local and global feature representation
Lim et al. (2020)	FgSegNet		√	Multi-scale-CNN, FPM	End-to-end foreground segmentation
Zhang et al. (2020)	Farsee-net		√	F-ASPP	Real-time multi-scale context aggregation
Gao et al. (2021)	MSCFNet		√	Multi-scale-CNN, EAR	Multi-scale contextual information
Hui et al. (2022)	Multi- branch Multi-scale-CNN		√	Multi-scale-CNN, DeepLabv3+	Adaptive multi-branch segmentation for small and obstructed targets

introduced a semantic mapping system to reconstruct dynamic vehicles and static roads. The fusion AR system first differentiates static and dynamic objects based on motion inconsistency (Rünz & Agapito, 2017), then tracks the six degrees of freedom pose of each model and aligns the pose with the previous frame. To improve coarse-grained semantic segmentation results, a network structure composed geometric segmentation and semantic segmentation (He et al., 2017) are applied. Some fusion AR systems train object labels associated with target models and fuse each object over time to achieve more accurate and robust performance (Keller et al., 2013; Whelan et al., 2015).

7 CONCLUSION

Although AR has been widely used to improve human experience in the virtual-reality intertwined world, the challenges of handling complex large-scale scenes, dealing with real-time occlusion, and rendering virtual objects have limited its applicability and performance. A potential solution approach to the challenges is a combination of image segmentation, vSLAM and machine learning. This paper investigates the recent advancements in AR visualization algorithms. Furthermore, this paper reviews deep learning-based vSLAMs, collaborative AR systems, and image segmentation-based AR systems to recommend the best solution approach for AR systems. Fusion systems are also introduced to

Figure 2. Basic structure of collaboration fusion AR systems



compensate for existing AR flaws. Fusion AR systems, which analyze contextual information through classification, detection, and tracking, will be a future research hotspot.

ACKNOWLEDGMENT

The authors would like to thank the Editor and the anonymous reviewers for their valuable comments.

FUNDING AGENCY

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

AlapattD.MascagniP.VardazaryanA.GarciaA.OkamotoN.MutterD.MarescauxJ.CostamagnaG.DallemagneB. PadoyN. (2021). Temporally Constrained Neural Networks (TCNN): A framework for semi-supervised video semantic segmentation.

Almalioglu, Y., Saputra, M. R. U., de Gusmao, P. P., Markham, A., & Trigoni, N. (2019). Ganvo: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks. In *International conference on robotics and automation,* (pp. 5474-5480). IEEE.

Alves, J., & Bernardino, A. (2020, April). A remote RGB-D VSLAM solution for low computational powered robots. In *2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC),* (pp. 214-220). IEEE. doi:10.1109/ICARSC49921.2020.9096074

Alzahrani, N. M., & Alfouzan, F. A. (2022). Augmented Reality (AR) and Cyber-Security for Smart Cities—A Systematic Literature Review. *Sensors (Basel)*, *22*(7), 2792. doi:10.3390/s22072792 PMID:35408406

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* (pp. 5297-5307). doi:10.1109/CVPR.2016.572

Arshad, S., & Kim, G. W. (2021). Role of deep learning in loop closure detection for visual and lidar SLAM: A survey. *Sensors (Basel)*, *21*(4), 1243. https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=33578695&dopt=Abstract doi:10.3390/s21041243 PMID:33578695

Azuma, R. T. (1997). A survey of augmented reality. *Presence (Cambridge, Mass.)*, *6*(4), 355–385. doi:10.1162/pres.1997.6.4.355

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(12), 2481–2495. https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=28060704&dopt=Abstract doi:10.1109/TPAMI.2016.2644615 PMID:28060704

Besl, P. J., & McKay, N. D. (1992). Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures* (Vol. 1611, pp. 586–606). International Society for Optics and Photonics. doi:10.1117/12.57955

Bian, X., Lim, S. N., & Zhou, N. (2016). *Multiscale fully convolutional network with application to industrial inspection. In 2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE.

Billinghurst, M., Clark, A., & Lee, G. (2015). *A survey of augmented reality*.

Bolya, D., Zhou, C., Xiao, F., & Lee, Y. J. (2020). Yolact++: Better real-time instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. PMID:32755851

Braud, T., Bijarbooneh, F. H., Chatzopoulos, D., & Hui, P. (2017). Future networking challenges: The case of mobile augmented reality. In *International Conference on Distributed Computing Systems,* (pp. 1796-1807). IEEE.

Caruso, D., Engel, J., & Cremers, D. (2015). Large-scale direct slam for omnidirectional cameras. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (pp. 141-148). IEEE. doi:10.1109/IROS.2015.7353366

Castle, R., Klein, G., & Murray, D. W. (2008). Video-rate localization in multiple maps for wearable augmented reality. In *International Symposium on Wearable Computers*, (pp. 15-22). IEEE. doi:10.1109/ISWC.2008.4911577

Castle, R. O., Gawley, D. J., Klein, G., & Murray, D. W. (2007). Towards simultaneous recognition, localization and mapping for hand-held and wearable cameras. In *Proceedings 2007 IEEE International Conference on Robotics and Automation,* (pp. 4102-4107). IEEE. doi:10.1109/ROBOT.2007.364109

ChenL. C.PapandreouG.KokkinosI.MurphyK.YuilleA. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848. https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=28463186&dopt=Abstract doi:10.1109/TPAMI.2017.2699184 PMID:28463186

Chen, S., Wu, J., Lu, Q., Wang, Y., & Lin, Z. (2021). Cross-scene loop-closure detection with continual learning for visual simultaneous localization and mapping. *International Journal of Advanced Robotic Systems*, *18*(5), 17298814211050560. doi:10.1177/17298814211050560

ChenZ.LamO.JacobsonA.MilfordM. (2014). Convolutional neural network-based place recognition.

Cheng, J., Sun, Y., & Meng, M. Q. H. (2019). Improving monocular visual SLAM in dynamic environments: An optical-flow-based approach. *Advanced Robotics*, *33*(12), 576–589. doi:10.1080/01691864.2019.1610060

Civera, J., Gálvez-López, D., Riazuelo, L., Tardós, J. D., & Montiel, J. M. M. (2011). Towards semantic SLAM using a monocular camera. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems,* (pp. 1277-1284). IEEE. doi:10.1109/IROS.2011.6094648

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* (pp. 3213-3223). doi:10.1109/CVPR.2016.350

Costa, G. D. M., Petry, M. R., & Moreira, A. P. (2022). Augmented Reality for Human–Robot Collaboration and Cooperation in Industrial Applications: A Systematic Literature Review. *Sensors (Basel)*, *22*(7), 2725. https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=35408339&dopt=Abstract doi:10.3390/s22072725 PMID:35408339

Costante, G., Mancini, M., Valigi, P., & Ciarfuglia, T. A. (2015). Exploring representation learning with cnns for frame-to-frame ego-motion estimation. *IEEE Robotics and Automation Letters*, *1*(1), 18–25. doi:10.1109/LRA.2015.2505717

Criminisi, A., Cross, G., Blake, A., & Kolmogorov, V. (2006). Bilayer segmentation of live video. In *Computer Society Conference on Computer Vision and Pattern Recognition,* 1, (pp. 53-60). IEEE.

Dai, W., Zhang, Y., Li, P., Fang, Z., & Scherer, S. (2020). Rgb-d slam in dynamic environments using point correlations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(1), 373–389. https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=32750826&dopt=Abstract doi:10.1109/TPAMI.2020.3010942 PMID:32750826

Davison, A. J. (2003). Real-time simultaneous localisation and mapping with a single camera. In *Computer Vision, 3*, (pp. 1403-1403). IEEE Computer Society. doi:10.1109/ICCV.2003.1238654

Davison, A. J., Reid, I. D., Molton, N. D., & Stasse, O. (2007). MonoSLAM: Real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6), 1052–1067. https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=17431302&dopt=Abstract doi:10.1109/TPAMI.2007.1049 PMID:17431302

Dey, A., Billinghurst, M., Lindeman, R. W., & Swan, J. II. (2018). A systematic review of 10 years of augmented reality usability studies: 2005 to 2014. *Frontiers in Robotics and AI*, *5*, 37. https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=33500923&dopt=Abstract doi:10.3389/frobt.2018.00037 PMID:33500923

Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S. R., Kowdle, A., & Izadi, S. (2016). Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics*, *35*(4), 1–13. doi:10.1145/2897824.2925969

Du, C., Chen, Y. L., Ye, M., & Ren, L. (2016). Edge snapping-based depth enhancement for dynamic occlusion handling in augmented reality. In *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR),* (pp. 54-62). IEEE. doi:10.1109/ISMAR.2016.17

Duan, C., Junginger, S., Huang, J., Jin, K., & Thurow, K. (2019). Deep learning for visual SLAM in transportation robotics: A review. *Transportation Safety and Environment*, *1*(3), 177–184. doi:10.1093/tse/tdz019

Eade, E., & Drummond, T. (2006). Edge Landmarks in Monocular SLAM. In BMVC, (pp. 7-16). doi:10.5244/C.20.2

Egger, J., & Masood, T. (2020). Augmented reality in support of intelligent manufacturing–a systematic literature review. *Computers & Industrial Engineering*, *140*, 106195. doi:10.1016/j.cie.2019.106195

Engel, J., Koltun, V., & Cremers, D. (2017). Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(3), 611–625. https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=28422651&dopt=Abstract doi:10.1109/TPAMI.2017.2658577 PMID:28422651

Engel, J., Schöps, T., & Cremers, D. (2014). LSD-SLAM: Large-scale direct monocular SLAM. In European conference on computer vision (pp. 834-849). Springer, Cham.

Engel, J., Stückler, J., & Cremers, D. (2015). Large-scale direct SLAM with stereo cameras. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 1935-1942). IEEE. doi:10.1109/IROS.2015.7353631

Engel, J., Sturm, J., & Cremers, D. (2013). Semi-dense visual odometry for a monocular camera. In *Proceedings of the IEEE international conference on computer vision* (pp. 1449-1456). doi:10.1109/ICCV.2013.183

Ess, A., Müller, T., Grabner, H., & Van Gool, L. (2009; Vol. 1). Segmentation-Based Urban Traffic Scene Understanding. In BMVC.

Forster, C., Lynen, S., Kneip, L., & Scaramuzza, D. (2013). Collaborative monocular slam with multiple micro aerial vehicles. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 3962-3970). IEEE. doi:10.1109/IROS.2013.6696923

Forster, C., Pizzoli, M., & Scaramuzza, D. (2014). *SVO: Fast semi-direct monocular visual odometry. In 2014 IEEE international conference on robotics and automation (ICRA)*. IEEE.

Fukiage, T., Oishi, T., & Ikeuchi, K. (2014). Visibility-based blending for real-time applications. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 63-72). IEEE. doi:10.1109/ISMAR.2014.6948410

Gao, G., Xu, G., Yu, Y., Xie, J., Yang, J., & Yue, D. (2021). MSCFNet: A lightweight network with multi-scale context fusion for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 1–11. doi:10.1109/TITS.2021.3098355

Gao, X., & Zhang, T. (2015). Loop closure detection for visual slam systems using deep neural networks. In 2015 34th Chinese Control Conference (CCC) (pp. 5851-5856). IEEE. doi:10.1109/ChiCC.2015.7260555

Gao, X., & Zhang, T. (2017). Unsupervised learning to detect loops using deep neural networks for visual SLAM system. *Autonomous Robots*, *41*(1), 1–18. doi:10.1007/s10514-015-9516-2

Gattullo, M., Evangelista, A., Uva, A. E., Fiorentino, M., & Gabbard, J. L. (2020). What, how, and why are visual assets used in industrial augmented reality? A systematic review and classification in maintenance, assembly, and training (from 1997 to 2019). *IEEE Transactions on Visualization and Computer Graphics*, *28*(2), 1443–1456. doi:10.1109/TVCG.2020.3014614 PMID:32759085

Geiger, A., Lenz, P., & Urtasun, R. (2012). *Are we ready for autonomous driving? the kitti vision benchmark suite. In 2012 IEEE conference on computer vision and pattern recognition*. IEEE.

Geng, J. (2011). Structured-light 3D surface imaging: A tutorial. *Advances in Optics and Photonics*, *3*(2), 128–160. doi:10.1364/AOP.3.000128

Goh, E. S., Sunar, M. S., & Ismail, A. W. (2019). 3D object manipulation techniques in handheld mobile augmented reality interface: A review. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 40581–40601. doi:10.1109/ACCESS.2019.2906394

Handa, A., Bloesch, M., Pătrăucean, V., Stent, S., McCormac, J., & Davison, A. (2016). gvnn: Neural network library for geometric computer vision. In *European Conference on Computer Vision* (pp. 67-82). Springer, Cham. doi:10.1007/978-3-319-49409-8_9

Hariharan, B., Arbeláez, P., Girshick, R., & Malik, J. (2014). Simultaneous detection and segmentation. In *European conference on computer vision* (pp. 297-312). Springer, Cham.

Hasinoff, S. W., Kang, S. B., & Szeliski, R. (2006). Boundary matting for view synthesis. *Computer Vision and Image Understanding*, *103*(1), 22–32. doi:10.1016/j.cviu.2006.02.005

He, H., Yuan, Y., Yue, X., & Hu, H. (2022). MLSeg: Image and video segmentation as multi-label classification and selected-label pixel classification. arXiv preprint arXiv:2203.04187.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961-2969).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Hebborn, A. K., Höhner, N., & Müller, S. (2017). Occlusion matting: realistic occlusion handling for augmented reality applications. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 62-71). IEEE. doi:10.1109/ISMAR.2017.23

Hentout, A., Maoudj, A., Kaid-Youcef, N., Hebib, D., & Bouzouia, B. (2020). Distributed multi-agent bidding-based approach for the collaborative mapping of unknown indoor environments by a homogeneous mobile robot team. *Journal of Intelligent Systems*, *29*(1), 84–99. doi:10.1515/jisys-2017-0255

Hirose, K., & Saito, H. (2012). Fast line description for line-based slam. In 2012 23rd British Machine Vision Conference, BMVC 2012. doi:10.5244/C.26.83

Hoff, W. A., Nguyen, K., & Lyon, T. (1996). Computer-vision-based registration techniques for augmented reality. In Intelligent Robots and Computer Vision XV: Algorithms, Techniques, Active Vision, and Materials Handling (Vol. 2904, pp. 538-548). International Society for Optics and Photonics.

Hou, J., Dai, A., & Nießner, M. (2019). 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4421-4430). doi:10.1109/CVPR.2019.00455

Hou, Y., Zhang, H., & Zhou, S. (2015). *Convolutional neural network-based image representation for visual loop closure detection. In 2015 IEEE international conference on information and automation*. IEEE.

Hu, P., Heilbron, F., Wang, O., Lin, Z., Sclaroff, S., & Perazzi, F. (2020). Temporally distributed networks for fast video semantic segmentation. arXiv preprint arXiv:2004.01800. 10.1109/CVPR42600.2020.00884

Huang, J., & You, S. (2016) Point Cloud Labeling using 3D Convolutional Neural Network. In 2016 22th International Conference on Pattern Recognition (ICPR). IEEE.

Huang, P., Zeng, L., Luo, K., Guo, J., Zhou, Z., & Chen, X. (2021, July). ColaSLAM: Real-Time Multi-Robot Collaborative Laser SLAM via Edge Computing. In *2021 IEEE/CIC International Conference on Communications in China (ICCC)*(pp. 242-247). IEEE. doi:10.1109/ICCC52777.2021.9580413

Huang, Z., Hui, P., Peylo, C., & Chatzopoulos, D. (2013). Mobile augmented reality survey: a bottom-up approach. arXiv preprint arXiv:1309.4413.

Hui, J., & Zhang, H. (2022). A semantic segmentation network based on multi-branch structures and multi-scale modules.

Innmann, M., Zollhöfer, M., Nießner, M., Theobalt, C., & Stamminger, M. (2016). Volumedeform: Real-time volumetric non-rigid reconstruction. In *European Conference on Computer Vision* (pp. 362-379). Springer, Cham.

Irshad, S., & Rambli, D. R. A. (2017, November). Advances in mobile augmented reality from user experience perspective: a review of studies. In *International Visual Informatics Conference* (pp. 466-477). Springer, Cham. doi:10.1007/978-3-319-70010-6_43

Irshad, S., & Rambli, D. R. B. A. (2014, September). User experience of mobile augmented reality: A review of studies. In *2014 3rd international conference on user science and engineering (i-USEr)* (pp. 125-130). IEEE. doi:10.1109/IUSER.2014.7002689

Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial transformer networks. *Advances in Neural Information Processing Systems*, *28*, 2017–2025.

Jang, Y., Oh, C., Lee, Y., & Kim, H. J. (2021). Multirobot collaborative monocular SLAM utilizing rendezvous. *IEEE Transactions on Robotics*, *37*(5), 1469–1486. doi:10.1109/TRO.2021.3058502

Ji, J., Shi, R., Li, S., Chen, P., & Miao, Q. (2020). Encoder-decoder with cascaded CRFs for semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(5), 1926–1938. doi:10.1109/TCSVT.2020.3015866

Jiao, L., Wang, D., Bai, Y., Chen, P., & Liu, F. (2021). Deep Learning in Visual Tracking: A Review. *IEEE Transactions on Neural Networks and Learning Systems*, 1–20. doi:10.1109/TNNLS.2021.3136907 PMID:34968181

Kähler, O., Prisacariu, V. A., Ren, C. Y., Sun, X., Torr, P., & Murray, D. (2015). Very high frame rate volumetric integration of depth images on mobile devices. *IEEE Transactions on Visualization and Computer Graphics*, *21*(11), 1241–1250. https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=26439825&dopt=Abstract doi:10.1109/TVCG.2015.2459891 PMID:26439825

Kaiming, H., Georgia, G., Piotr, D., & Ross, G. S. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961-2969).

Kakuta, T., Vinh, L. B., Kawakami, R., Oishi, T., & Ikeuchi, K. (2008). Detection of moving objects and cast shadows using a spherical vision camera for outdoor mixed reality. In *Proceedings of the 2008 ACM symposium on Virtual reality software and technology* (pp. 219-222). doi:10.1145/1450579.1450626

Kalkofen, D., Sandor, C., White, S., & Schmalstieg, D. (2011). Visualization techniques for augmented reality. In *Handbook of augmented reality* (pp. 65–98). Springer. doi:10.1007/978-1-4614-0064-6_3

Kalogerakis, E., Averkiou, M., Maji, S., & Chaudhuri, S. (2017). 3D shape segmentation with projective convolutional networks. In proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3779-3788).

Kanbara, M., Okuma, T., Takemura, H., & Yokoya, N. (1999). Real-time composition of stereo images for video see-through augmented reality. In *Proceedings IEEE International Conference on Multimedia Computing and Systems* (Vol. 1, pp. 213-219). IEEE. doi:10.1109/MMCS.1999.779195

Karrer, M., Schmuck, P., & Chli, M. (2018). CVI-SLAM—Collaborative visual-inertial SLAM. *IEEE Robotics and Automation Letters*, *3*(4), 2762–2769. doi:10.1109/LRA.2018.2837226

Kaygusuz, N., Mendez, O., & Bowden, R. (2021, September). Multi-Camera Sensor Fusion for Visual Odometry using Deep Uncertainty Estimation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)* (pp. 2944-2949). IEEE. doi:10.1109/ITSC48978.2021.9565079

Keivan, N., Patron-Perez, A., & Sibley, G. (2016). Asynchronous adaptive conditioning for visual-inertial SLAM. In *Experimental Robotics* (pp. 309–321). Springer. doi:10.1007/978-3-319-23778-7_21

Keller, M., Lefloch, D., Lambers, M., Izadi, S., Weyrich, T., & Kolb, A. (2013). Real-time 3d reconstruction in dynamic scenes using point-based fusion. In 2013 International Conference on 3D Vision-3DV 2013 (pp. 1-8). IEEE.

Kim, H., Yang, S. J., & Sohn, K. (2003). 3d reconstruction of stereo images for interaction between real and virtual worlds. In The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings. (pp. 169-176). IEEE.

Kim, T. H., Jung, H., Lee, K. M., & Lee, S. U. (2008). Segment-based foreground object disparity estimation using Zcam and multiple-view stereo. In *2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (pp. 1251-1254). IEEE. doi:10.1109/IIH-MSP.2008.343

Kirillov, A., He, K., Girshick, R., Rother, C., & Dollár, P. (2019). Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9404-9413).

Kiss-Illés, D., Barrado, C., & Salamí, E. (2019). GPS-SLAM: An augmentation of the ORB-SLAM algorithm. *Sensors (Basel)*, *19*(22), 4973. doi:10.3390/s19224973 PMID:31731624

Klein, G., & Murray, D. (2007). Parallel tracking and mapping for small AR workspaces. In 2007 6th IEEE and ACM international symposium on mixed and augmented reality (pp. 225-234). IEEE. doi:10.1109/ISMAR.2007.4538852

Klein, G., & Murray, D. (2007). Parallel tracking and mapping for small AR workspaces. In 2007 6th IEEE and ACM international symposium on mixed and augmented reality (pp. 225-234). IEEE. doi:10.1109/ISMAR.2007.4538852

Klette, R., Koschan, A., & Schluns, K. (1998). *Three-dimensional data from images*. Springer-Verlag Singapore Pte. Ltd.

Koh, Y. S., Goh, K. W., Dares, M., Yeong, C. F., Ming, E. S. L., Sunar, M. S., & Tey, Y. S. (2020). A review on augmented reality tracking methods for maintenance of robots. *Jurnal Teknologi*, *83*(1), 37–43. doi:10.11113/jurnalteknologi.v83.14907

Konda, K. R., & Memisevic, R. (2015). Learning visual odometry with a convolutional network. In VISAPP (1), (pp. 486-490). doi:10.5220/0005299304860490

Krähenbühl, P., & Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in Neural Information Processing Systems*, *24*, 109–117.

Krähenbühl, P., & Koltun, V. (2013). Parameter learning and convergent inference for dense random fields. In *International Conference on Machine Learning* (pp. 513-521). PMLR.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Lee, L. H., Braud, T., Hosio, S., & Hui, P. (2021). Towards Augmented Reality Driven Human-City Interaction: Current Research on Mobile Headsets and Future Challenges. [CSUR]. *ACM Computing Surveys*, *54*(8), 1–38.

Li, P., Wang, D., Wang, L., & Lu, H. (2018). Deep visual tracking: Review and experimental comparison. *Pattern Recognition*, *76*, 323–338. doi:10.1016/j.patcog.2017.11.007

Li, X., Yi, W., Chi, H. L., Wang, X., & Chan, A. P. (2018). A critical review of virtual and augmented reality (VR/AR) applications in construction safety. *Automation in Construction*, *86*, 150–162. doi:10.1016/j.autcon.2017.11.003

Li, X., Zhang, L., & Zhu, Z. (2022). SnapshotNet: Self-supervised feature learning for point cloud data segmentation using minimal labeled data. *Computer Vision and Image Understanding*, *216*, 103339. doi:10.1016/j.cviu.2021.103339

Li, Y., Shi, J., & Lin, D. (2018). Low-latency video semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5997-6005).

Li, Y., Sun, J., Tang, C. K., & Shum, H. Y. (2004). Lazy snapping. *ACM Transactions on Graphics*, *23*(3), 303–308. doi:10.1145/1015706.1015719

Lim, L. A., & Keles, H. Y. (2020). Learning multi-scale features for foreground segmentation. *Pattern Analysis & Applications*, *23*(3), 1369–1380. doi:10.1007/s10044-019-00845-9

Lin, D., Ji, Y., Lischinski, D., Cohen-Or, D., & Huang, H. (2018). Multi-scale context intertwining for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 603-619).

Ling, H. (2017). Augmented reality in reality. *IEEE MultiMedia*, *24*(3), 10–15. doi:10.1109/MMUL.2017.3051517

Liu, R., Yang, J., Chen, Y., & Zhao, W. (2019, June). eslam: An energy-efficient accelerator for real-time orb-slam on fpga platform. In *Proceedings of the 56th Annual Design Automation Conference 2019* (pp. 1-6). doi:10.1145/3316781.3317820

Liu, Z., Suo, C., Zhou, S., Xu, F., Wei, H., Chen, W., & Liu, Y. H. et al. (2019, November). Seqlpd: Sequence matching enhanced loop-closure detection based on large-scale point cloud description for self-driving vehicles. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 1218-1223). IEEE. doi:10.1109/IROS40897.2019.8967875

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

Masood, T., & Egger, J. (2019). Augmented reality in support of Industry 4.0—Implementation challenges and success factors. *Robotics and Computer-integrated Manufacturing*, *58*, 181–195. doi:10.1016/j.rcim.2019.02.003

McCormac, J., Handa, A., Davison, A., & Leutenegger, S. (2017). Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and automation (ICRA)* (pp. 4628-4635). IEEE. doi:10.1109/ICRA.2017.7989538

Mei, C., Sibley, G., Cummins, M., Newman, P. M., & Reid, I. (2009). A Constant-Time Efficient Stereo SLAM System. In BMVC (pp. 1-11). doi:10.5244/C.23.54

Memon, A. R., Wang, H., & Hussain, A. (2020). Loop closure detection using supervised and unsupervised deep neural networks for monocular SLAM systems. *Robotics and Autonomous Systems*, *126*, 103470. doi:10.1016/j.robot.2020.103470

Merrill, N., & Huang, G. (2019, November). CALC2. 0: Combining appearance, semantic and geometric information for robust and efficient visual loop closure. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 4554-4561). IEEE. doi:10.1109/IROS40897.2019.8968159

Mur-Artal, R., & Tardós, J. D. (2014). ORB-SLAM: tracking and mapping recognizable features. In Workshop on Multi View Geometry in Robotics (MVIGRO)-RSS (Vol. 2014, p. 2).

Mur-Artal, R., & Tardós, J. D. (2017). Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, *33*(5), 1255–1262. doi:10.1109/TRO.2017.2705103

Newcombe, R. A., Fox, D., & Seitz, S. M. (2015). Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* (pp. 343-352). doi:10.1109/CVPR.2015.7298631

Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., . . . Fitzgibbon, A. (2011). Kinectfusion: Real-time dense surface mapping and tracking. In *International symposium on mixed and augmented reality*, (pp. 127-136). IEEE.

Newcombe, R. A., Lovegrove, S. J., & Davison, A. J. (2011). *DTAM: Dense tracking and mapping in real-time. In 2011 international conference on computer vision*. IEEE.

Nistér, D. (2004). An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(6), 756–770. doi:10.1109/TPAMI.2004.17 PMID:18579936

Nistér, D., & Stewénius, H. (2007). A minimal solution to the generalised 3-point pose problem. *Journal of Mathematical Imaging and Vision*, *27*(1), 67–79. doi:10.1007/s10851-006-0450-y

OberwegerM.WohlhartP.LepetitV. (2015). Hands deep in deep learning for hand pose estimation.

Okutomi, M., & Kanade, T. (1993). A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(4), 353–363. doi:10.1109/34.206955

Ondrúška, P., Kohli, P., & Izadi, S. (2015). Mobilefusion: Real-time volumetric surface reconstruction and dense tracking on mobile phones. *IEEE Transactions on Visualization and Computer Graphics*, *21*(11), 1251–1258. doi:10.1109/TVCG.2015.2459902 PMID:26439826

Outahar, M., Moreau, G., & Normand, J. M. (2021). Direct and Indirect vSLAM Fusion for Augmented Reality. *Journal of Imaging*, *7*(8), 141. https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=34460777&dopt=Abstract doi:10.3390/jimaging7080141 PMID:34460777

Palmarini, R., Erkoyuncu, J. A., Roy, R., & Torabmostaedi, H. (2018). A systematic review of augmented reality applications in maintenance. *Robotics and Computer-integrated Manufacturing*, *49*, 215–228. doi:10.1016/j.rcim.2017.06.002

PaszkeA.ChaurasiaA.KimS.CulurcielloE. (2016). Enet: A deep neural network architecture for real-time semantic segmentation.

Paul, M., Mayer, C., Gool, L., & Timofte, R. (2020) Efficient video semantic segmentation with labels propagation and refinement. In *Winter Conference on Applications of Computer Vision (WACV)* (pp. 2873-2882). IEEE. doi:10.1109/WACV45572.2020.9093520

Perron, J. M., Huang, R., Thomas, J., Zhang, L., Tan, P., & Vaughan, R. T. (2015). Orbiting a moving target with multi-robot collaborative visual slam. In *Workshop on Multi-View Geometry in Robotics (MVIGRO),* (pp. 1339-1344).

PinheiroP. O.CollobertR.DollárP. (2015). Learning to segment object candidates.

Qi, C. R., Liu, W., Wu, C., Su, H., & Guibas, L. J. (2018). Frustum point nets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* (pp. 918-927).

Qin, X., Wang, B., Boegner, D., Gaitan, B., Zheng, Y., Du, X., & Chen, Y. (2021). Indoor localization of hand-held OCT probe using visual odometry and real-time segmentation using deep learning. *IEEE Transactions on Biomedical Engineering*, *69*(4), 1378–1385. doi:10.1109/TBME.2021.3116514 PMID:34587002

Qiu, K., Ai, Y., Tian, B., Wang, B., & Cao, D. (2018). Siamese-ResNet: implementing loop closure detection based on siamese network. In *2018 IEEE Intelligent Vehicles Symposium (IV),* (pp. 716-721). IEEE. doi:10.1109/IVS.2018.8500465

Rabbi, I., & Ullah, S. (2013). A survey on augmented reality challenges and tracking. *Acta graphica: znanstveni časopis za tiskarstvo i grafičke komunikacije, 24*(1-2), 29-46.

Rabbi, I., Ullah, S., & Khan, S. U. (2012). Augmented reality tracking techniques—A systematic literature. *IOSR Journal of Computer Engineering*, *2*(2), 23–29. doi:10.9790/0661-0222329

Raj, A., Maturana, D., & Scherer, S. (2015). Multi-scale convolutional architecture for semantic segmentation. Robotics Institute, Carnegie Mellon University, Tech. Rep. CMU-RITR-15-21.

Riazuelo, L., Civera, J., & Montiel, J. M. (2014). C2tam: A cloud framework for cooperative tracking and mapping. *Robotics and Autonomous Systems*, *62*(4), 401–413. doi:10.1016/j.robot.2013.11.007

Rother, C., Kolmogorov, V., & Blake, A. (2004). " GrabCut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG), 23*(3), 309-314.

Roxas, M., Hori, T., Fukiage, T., Okamoto, Y., & Oishi, T. (2018). Occlusion handling using semantic segmentation and visibility-based rendering for mixed reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology,* (pp. 1-8). doi:10.1145/3281505.3281546

Roy, A., & Todorovic, S. (2016). A multi-scale cnn for affordance segmentation in rgb images. In *European conference on computer vision* (pp. 186-201). Springer, Cham. doi:10.1007/978-3-319-46493-0_12

Rudin, L. I., Osher, S., & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D. Nonlinear Phenomena*, *60*(1-4), 259–268. doi:10.1016/0167-2789(92)90242-F

Rünz, M., & Agapito, L. (2017). Co-fusion: Real-time segmentation, tracking and fusion of multiple objects. In *International Conference on Robotics and Automation (ICRA)*, (pp. 4471-4478). IEEE. doi:10.1109/ICRA.2017.7989518

Runz, M., Buffier, M., & Agapito, L. (2018). Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In *International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 10-20). IEEE. doi:10.1109/ISMAR.2018.00024

Salas-Moreno, R. F., Glocken, B., Kelly, P. H., & Davison, A. J. (2014). *Dense planar SLAM. In 2014 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE.

Salas-Moreno, R. F., Newcombe, R. A., Strasdat, H., Kelly, P. H., & Davison, A. J. (2013). Slam++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1352-1359). doi:10.1109/CVPR.2013.178

Schmuck, P., & Chli, M. (2019). CCM-SLAM: Robust and efficient centralized collaborative monocular simultaneous localization and mapping for robotic teams. *Journal of Field Robotics*, *36*(4), 763–781. doi:10.1002/rob.21854

Schmuck, P., Ziegler, T., Karrer, M., Perraudin, J., & Chli, M. (2021). COVINS: Visual-Inertial SLAM for Centralized Collaboration. In *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (pp. 171-176). IEEE. doi:10.1109/ISMAR-Adjunct54149.2021.00043

Schöps, T., Engel, J., & Cremers, D. (2014). *Semi-dense visual odometry for AR on a smartphone. In 2014 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE.

Shafi, M., Molisch, A. F., Smith, P. J., Haustein, T., Zhu, P., De Silva, P., Tufvesson, F., Benjebbour, A., & Wunder, G. (2017). 5G: A tutorial overview of standards, trials, challenges, deployment, and practice. *IEEE Journal on Selected Areas in Communications*, *35*(6), 1201–1221. doi:10.1109/JSAC.2017.2692307

Shelhamer, E., Rakelly, K., Hoffman, J., & Darrell, T. (2016). Clockwork convnets for video semantic segmentation. In *European Conference on Computer Vision,* (pp. 852-868). Springer, Cham.

Shotton, J., Winn, J., Rother, C., & Criminisi, A. (2009). Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, *81*(1), 2–23. doi:10.1007/s11263-007-0109-1

SimonyanK.ZissermanA. (2014). Very deep convolutional networks for large-scale image recognition.

SM, & Augasta, G. M. (2021). Review of recent advances in visual tracking techniques. *Multimedia Tools and Applications*, *80*(16), 24185–24203. doi:10.1007/s11042-021-10848-6

Spittle, B., Frutos-Pascual, M., Creed, C., & Williams, I. (2022). A Review of Interaction Techniques for Immersive Environments. *IEEE Transactions on Visualization and Computer Graphics*, 1. https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=35552136&dopt=Abstract doi:10.1109/TVCG.2022.3174805 PMID:35552136

Strasdat, H., Montiel, J. M., & Davison, A. J. (2012). Visual SLAM: Why filter? *Image and Vision Computing*, *30*(2), 65–77. doi:10.1016/j.imavis.2012.02.009

Stühmer, J., Gumhold, S., & Cremers, D. (2010). Real-time dense geometry from a handheld camera. In *Joint Pattern Recognition Symposium,* (pp. 11-20). Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-15986-2_2

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

Taketomi, T., Uchiyama, H., & Ikeda, S. (2017). Visual SLAM algorithms: A survey from 2010 to 2016. *IPSJ Transactions on Computer Vision and Applications*, *9*(1), 1–11. doi:10.1186/s41074-017-0027-2

Tang, X., Hu, X., Fu, C. W., & Cohen-Or, D. (2020). GrabAR: Occlusion-aware Grabbing Virtual Objects in AR. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, (pp. 697-708). doi:10.1145/3379337.3415835

Tateno, K., Tombari, F., Laina, I., & Navab, N. (2017). Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 6243-6252). doi:10.1109/CVPR.2017.695

Tateno, K., Tombari, F., & Navab, N. (2016). *When 2.5 D is not enough: Simultaneous reconstruction, segmentation and recognition on dense SLAM. In 2016 IEEE international conference on robotics and automation (ICRA)*. IEEE.

Tian, Z., Shen, C., Wang, X., & Chen, H. (2021). Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* (pp. 5443-5452). doi:10.1109/CVPR46437.2021.00540

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2016). Deep end2end voxel2voxel prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops,* (pp. 17-24).

Triputen, S., Gopal, A., Weber, T., Höfert, C., Rätsch, M., & Schreve, K. (2018, March). Methodology to analyze the accuracy of 3D objects reconstructed with collaborative robot based monocular LSD-SLAM. In *2018 International Conference on Intelligent Autonomous Systems (Icoias),* (pp. 185-190). IEEE. doi:10.1109/ICoIAS.2018.8494109

Uhrig, J., Rehder, E., Fröhlich, B., Franke, U., & Brox, T. (2018, June). Box2pix: Single-shot instance segmentation by assigning pixels to object boxes. In *2018 IEEE Intelligent Vehicles Symposium (IV)* (pp. 292-299). IEEE. doi:10.1109/IVS.2018.8500621

Van Krevelen, D. W. F., & Poelman, R. (2010). A survey of augmented reality technologies, applications and limitations. *The International Journal of Virtual Reality: a Multimedia Publication for Professionals*, *9*(2), 1–20. doi:10.20870/IJVR.2010.9.2.2767

Van Opdenbosch, D., & Steinbach, E. (2018). Collaborative visual slam using compressed feature exchange. *IEEE Robotics and Automation Letters*, *4*(1), 57–64. doi:10.1109/LRA.2018.2878920

WangH.WangW.LiuJ. (2021) Temporal memory attention for video semantic segmentation. 10.1109/ICIP42928.2021.9506731

Wang, K., Ma, S., Chen, J., Ren, F., & Lu, J. (2020). Approaches challenges and applications for deep visual odometry toward to complicated and emerging areas. *IEEE Transactions on Cognitive and Developmental Systems*.

Wang, S., Clark, R., Wen, H., & Trigoni, N. (2018). End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks. *The International Journal of Robotics Research*, *37*(4-5), 513–542. doi:10.1177/0278364917734298

Wang, X., Zhang, R., Kong, T., Li, L., & Shen, C. (2020). Solov2: Dynamic and fast instance segmentation. *Advances in Neural Information Processing Systems*, *33*, 17721–17732.

Wang, Y., Wang, P., Luo, Z., & Yan, Y. (2022). A novel AR remote collaborative platform for sharing 2.5 D gestures and gaze. *International Journal of Advanced Manufacturing Technology*, 1–9. PMID:35095164

Westphal, C. (2017). *Challenges in networking to support augmented reality and virtual reality*. IEEE ICNC.

Whelan, T., Leutenegger, S., Salas-Moreno, R., Glocker, B., & Davison, A. (2015). ElasticFusion: Dense SLAM without a pose graph. *Robotics Science and Systems: Online Proceedings*. doi:10.15607/RSS.2015.XI.001

Williams, B., Klein, G., & Reid, I. (2007). Real-time SLAM relocalisation. In *international conference on computer vision,* (pp. 1-8). IEEE. doi:10.1109/ICCV.2007.4409115

Wu, B., Zhou, X., Zhao, S., Yue, X., & Keutzer, K. (2019). Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *2019 International Conference on Robotics and Automation (ICRA),* (pp. 4376-4382). IEEE. doi:10.1109/ICRA.2019.8793495

Xu, J., Cao, H., Yang, Z., Shangguan, L., Zhang, J., He, X., & Liu, Y. (2022). {SwarmMap}: Scaling Up Real-time Collaborative Visual {SLAM} at the Edge. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)* (pp. 977-993).

Yao, E., Zhang, H., Xu, H., Song, H., & Zhang, G. (2018). Robust RGB-D visual odometry based on edges and points. *Robotics and Autonomous Systems*, *107*, 209–220. doi:10.1016/j.robot.2018.06.009

YuF.KoltunV. (2015). Multi-scale context aggregation by dilated convolutions.

ZagoruykoS.LererA.LinT. Y.PinheiroP. O.GrossS.ChintalaS.DollárP. (2016). A multipath network for object detection. 10.5244/C.30.15

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision,* (pp. 818-833). Springer, Cham.

Zeiler, M. D., Taylor, G. W., & Fergus, R. (2011). Adaptive deconvolutional networks for mid and high level feature learning. In *International Conference on Computer Vision,* (pp. 2018-2025). IEEE. doi:10.1109/ICCV.2011.6126474

Zhang, H., Chen, X., Lu, H., & Xiao, J. (2018). Distributed and collaborative monocular simultaneous localization and mapping for multi-robot systems in large-scale environments. *International Journal of Advanced Robotic Systems*, *15*(3), 1729881418780178. doi:10.1177/1729881418780178

Zhang, H., Jiang, K., Zhang, Y., Li, Q., Xia, C., & Chen, X. (2014). Discriminative feature learning for video semantic segmentation. In *2014 International Conference on Virtual Reality and Visualization* (pp. 321-326). IEEE. doi:10.1109/ICVRV.2014.65

Zhang, H., Wang, K., Tian, Y., Gou, C., & Wang, F. Y. (2018). MFR-CNN: Incorporating multi-scale features and global information for traffic object detection. *IEEE Transactions on Vehicular Technology*, *67*(9), 8019–8030. doi:10.1109/TVT.2018.2843394

Zhang, L., Wang, L., Zhang, X., Shen, P., Bennamoun, M., Zhu, G., Shah, S. A. A., & Song, J. (2018). Semantic scene completion with dense CRF from a single depth image. *Neurocomputing*, *318*, 182–195. doi:10.1016/j.neucom.2018.08.052

Zhang, S., Lu, S., He, R., & Bao, Z. (2021). Stereo Visual Odometry Pose Correction through Unsupervised Deep Learning. *Sensors (Basel)*, *21*(14), 4735. doi:10.3390/s21144735 PMID:34300475

Zhang, T., Wei, S., & Ji, S. (2022). E2EC: An End-to-End Contour-based Method for High-Quality High-Speed Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4443-4452).

Zhang, Z. (2012). Microsoft kinect sensor and its effect. *IEEE MultiMedia*, *19*(2), 4–10. doi:10.1109/MMUL.2012.24

Zhang, Z., & Zhang, K. (2020, May). Farsee-net: Real-time semantic segmentation by efficient multi-scale context aggregation and feature space super-resolution. In *2020 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 8411-8417). IEEE. doi:10.1109/ICRA40945.2020.9196599

Zhao, H., Qi, X., Shen, X., Shi, J., & Jia, J. (2018). Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 405-420). doi:10.1007/978-3-030-01219-9_25

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., & Torr, P. H. et al. (2015). Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision,* (pp. 1529-1537). doi:10.1109/ICCV.2015.179

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 633-641).

Zhu, J., Wang, L., Yang, R., & Davis, J. (2008). Fusion of time-of-flight depth and stereo for high accuracy depth maps. In *Conference on Computer Vision and Pattern Recognition,* (pp. 1-8). IEEE.

ZhuX. F.XuT.WuX. J. (2022). Visual Object Tracking on Multi-modal RGB-D Videos: A Review.

Zollhöfer, M., Nießner, M., Izadi, S., Rehmann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., Loop, C., Theobalt, C., & Stamminger, M. (2014). Real-time non-rigid reconstruction using an RGB-D camera. *ACM Transactions on Graphics*, *33*(4), 1–12. doi:10.1145/2601097.2601165

Zou, D., & Tan, P. (2012). Coslam: Collaborative visual slam in dynamic environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(2), 354–366. doi:10.1109/TPAMI.2012.104 PMID:22547430

Zou, D., Tan, P., & Yu, W. (2019). Collaborative visual SLAM for multiple agents: A brief survey. *Virtual Reality & Intelligent Hardware*, *1*(5), 461–482. doi:10.1016/j.vrih.2019.09.002