Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

Deep learning-based object detection in augmented reality: A systematic review

Yalda Ghasemi[a], Heejin Jeong[a,*], Sung Ho Choi[b], Kyeong-Beom Park[b], Jae Yeol Lee[b,*]

[a] *Department of Mechanical and Industrial Engineering, University of Illinois at Chicago, USA*
[b] *Department of Industrial Engineering, Chonnam National University, South Korea*

ARTICLE INFO

Article history:
Received 21 December 2021
Received in revised form 18 March 2022
Accepted 19 March 2022
Available online 6 April 2022

Keywords:
Deep learning
Object detection
Augmented reality
Mixed reality
Computation platform
Systematic review

ABSTRACT

Recent advances in augmented reality (AR) and artificial intelligence have caused these technologies to pioneer innovation and alteration in any field and industry. The fast-paced developments in computer vision (CV) and augmented reality facilitated analyzing and understanding the surrounding environments. This paper systematically reviews and presents studies that integrated augmented/mixed reality and deep learning for object detection over the past decade. Five sources including Scopus, Web of Science, IEEE Xplore, ScienceDirect, and ACM were used to collect data. Finally, a total of sixty-nine papers were analyzed from two perspectives: (1) application analysis of deep learning-based object detection in the context of augmented reality and (2) analyzing the use of servers or local AR devices to perform the object detection computations to understand the relation between object detection algorithms and AR technology. Furthermore, the advantages of using deep learning-based object detection to solve the AR problems and limitations hindering the ultimate use of this technology are critically discussed. Our findings affirm the promising future of integrating AR and CV.

Contents

[*] Corresponding authors.
E-mail addresses: heejinj@uic.edu (H. Jeong), jaeyeol@chonnam.ac.kr (J.Y. Lee).

https://doi.org/10.1016/j.compind.2022.103661
0166-3615/© 2022 Elsevier B.V. All rights reserved.

1. Introduction

The advent of the digitalized world has enhanced the quality of information and resulted in a tremendous amount of data generation. A wide variety of technologies leverage these data to provide novel solutions and perform traditional tasks more innovatively. Augmented reality (AR) is a prominent example of such technologies, which has become one of the most popular technology trends in the current era (del Amo et al., 2018; Geng et al., 2020). AR can be defined as an extended version of the physical world overlaid with digital content bridging the real and virtual environments. AR systems should accurately identify the real environment and its components to work best. Most AR technologies can recognize a 3D spatial map of the components by real-time scanning of the environment.

There is a rapid growth of marker-based and markerless techniques to identify real-world content in AR systems (Katiyar et al., 2015). In marker-based AR, objects can be localized and tracked using physical markers attached to real objects. This technique often suffers from inaccurate detections, especially when there is a considerable distance between the AR camera and the real object or an obstacle between the AR camera and the real object causing occlusion. In addition, the markers should not reflect light, and their black and white colors should have a strong contrast. To overcome such limitations, markerless AR was proposed to recognize the real environment based on spatial geometry and does not require any trigger markers attached to the space. This technique often results in more accurate detections by obtaining the spatial map of the environment and its components. However, markerless applications need to recognize a textured flat surface to augment the digital content effectively.

Deep learning techniques can overcome typical marker-based and markerless AR deficiencies and provide faster and more accurate detections. The final goal of deep learning-based object detection is to recognize and locate one or multiple objects in a specific frame. The entire frame is the input, and the output is the object's bounding box and its probability of recognizing an object. Deep learning-based object detection identifies the existing objects in an image or video and demonstrates where they are located (i.e., object localization) and which category they belong to (i.e., object classification) (Sharma et al., 2020; He et al., 2016). A bounding box will discriminate the detected object from the background and other objects in the frame. Furthermore, image segmentation can be applied to assign particular class labels to each pixel of an image (Chen et al., 2014). The performance of image segmentation is hampered when one image includes multiple objects of the same class. To overcome this issue, instance segmentation was proposed to provide a pixel-level localization distinction among the objects by discerning between the objects of the same class.

Deep learning has been reviewed and investigated extensively for object detection applications. Previous studies used deep learning for object detection in various applications (Zamora-Hernández et al., 2021; Park et al., 2021; Khan et al., 2021). However, less attention has been paid to reviewing deep learning-based object detection techniques in AR and their current and future directions. This paper provides a comprehensive literature review and an in-depth discussion of deep learning-based object detection in AR and describes significant research trends in this field. In addition, this study elaborates on how these techniques can be used to enhance the performance of an AR system. It also examines the effectiveness of deep learning algorithms for object detection in AR, along with their advantages and disadvantages.

The remainder of this paper is organized as follows. In Section 2, the AR technology and its standard devices are introduced. Section 3 presents an introduction to deep learning-based object detection algorithms. In Section 4, the review method of this paper is described. In Section 5, the previous studies on object detection in AR are reviewed, and important information of each reviewed article is summarized in Section 6. Section 7 presents a discussion on the limitations of the current deep learning-based object detection approaches in AR. Finally, Section 8 includes conclusions and future directions of this area of research.

2. Augmented reality technologies and devices

This section provides an introduction to AR technologies and devices along with their advantages and disadvantages. AR was first introduced in the 1960s with limited functionalities. Still, it has made tremendous progress over the past decades and is becoming increasingly popular and being used in many industries and applications. Generally, many devices support AR technology. However, the capabilities may differ to some degree from one to another. Despite this difference, all AR devices consist of two elements: an image generating optical unit producing the virtual content, and a projection surface displaying the produced virtual content to the users. AR devices can generally be divided into four categories: wearable devices, handheld devices, projection-based displays (also known as spatial AR), and holographic displays.

2.1. Wearable devices

Wearable AR devices are the most advanced devices typically worn on the user's head. These devices are often used as helmets, goggles, or glasses. Wearables are also called optical heads-up displays since they can superimpose computer-generated content into a see-through display in front of the user's eye. These devices do not impede the user's vision, and they are only responsible for augmenting the digital content to the scene. Wearables give the users the freedom to use both of their hands while interacting with AR content. Popular wearables (e.g., Microsoft HoloLens, Google glasses, Epson Moverio, and Magic Leap) are being used in the military to train soldiers in simulated environments (Livingston et al., 2011; Hidalgo et al., 2021), in the industry to provide real-time smart task assistance to the workers (Park et al., 2020a), or to provide data entry interfaces to the office workers (Singh et al., 2021). In addition, these devices are generally more unlikely to be accepted by the public, considering their current design.

2.2. Handheld devices (HHD)

Handheld devices (HHD) or mobile devices are one of the popular and easy-to-use devices for AR applications. Mobile AR allows users to create snapshots of the enhanced environments. Many camera-equipped devices such as smartphones or tablets can support AR and make the AR experience available. Since the acceptability of some AR devices, especially in public, is still a challenge, HHDs are a practical solution for the everyday use of AR technology, such as navigation and gaming. HHDs are lightweight and do not need sophisticated hardware requirements to offer an AR experience (Chowdhury et al., 2013).

2.3. Projector-based displays

Projectors directly overlay the information into the physical world without any mediator device. A projector-based display can turn any surface into a screen using the projection mapping technique responsible for scanning the environment by combining visible-light cameras with depth sensors to map the shape of the objects. It then overlays the 2D content ranging from images, videos, or only guidance lights onto the environment. Users have the freedom of working with two hands when working with elements created by projector-based displays, and the user does not need to wear a bulky headset. Furthermore, working with projector-based displays does not require an extensive amount of training. However, it suffers from some disadvantages. Since the projectors are not equipped with Inertial Measurement Unit (IMU) sensors, other external sensors should be used to make the interactions possible. Moreover, physical objects can occlude the AR information. In addition, since the projector is fixed to a particular location, all parts of the real environment may not be overlaid with AR content in complex physical environments.

2.4. Holographic displays

A holographic display is a form of display that creates 3D digital content using diffractions of light. Like projector-based displays, holographic displays do not need a device to mediate showing the augmented information to the user (Lin and Wu, 2017). Holographic displays use holograms instead of graphic images to produce projected pictures. They beam white light or laser light onto the holograms. The projected light produces bright two- or three-dimensional images. While plain daylight facilitates the creation of simple holograms, true 3-D images require laser-based holographic projectors. Such images can be viewed from different angles and a true perspective.

3. Deep learning-based object detection

This section provides an introduction of the most common algorithms used in deep learning-based object detection as well as their advantages and disadvantages. Object detection is important in computer vision since it intelligently identifies and analyzes a scene in a given frame. Depending on the context, the detection task can be divided into several categories, such as face detection, pose detection, and pedestrian detection, all of which have been used in various applications such as autonomous vehicles (Y. Li et al., 2020), robotics (Choi et al., 2022; Zhou et al., 2022; Chen et al., 2008), and security (Jain, 2019). Applying deep learning-based object detection in AR has been a challenge for researchers. The collected data is usually used to train a model if it is unique and the existing models are unsatisfactory. Otherwise, a pre-trained model is used to make the predictions. There are mainly two approaches to implement these computations, including server and local devices. Depending on the complexity of the model and the amount of the training data, the training process can happen either on a local device or on a remote server.

3.1. Convolutional neural network (CNN)

Convolutional neural networks or CNNs in short are the simplest and most widely used deep learning algorithms for object detection. In CNN, first, an image should be divided into separate pieces. The algorithm takes each of these pieces as the input and after going through convolutions and pooling layers, it outputs the objects' classes. The problem with this approach is that the objects may cover different frame ratios. Therefore, it requires many regions, which results in a massive amount of computational time. To overcome these issues, a faster approach is needed to reduce the number of regions by obtaining them through proposal methods.

3.2. Region-based convolutional neural network (R-CNN)

Region-based Convolutional Neural Network or R-CNN works on a specific number of regions. The algorithm extracts a group of boxes or Regions of Interest (ROI) using proposal methods such as Selective Search (Uijlings et al., 2013) in the frame and checks whether an object exists in that specific region. To use the R-CNN algorithm, first, it is required to choose a pre-trained convolutional neural network. Then, based on the number of classes that need to be detected, the network's last layer should be re-trained. Next, the regions should be reshaped to be matched to the network input size. After retrieving the regions, typical classifiers such as Support Vector Machine (SVM) can be used to classify the objects. Finally, techniques such as linear regression are used to assign a bounding box to each predicted class. Although R-CNN is a practical algorithm for object detection, it often suffers from low computational speed.

3.3. Fast R-CNN

Fast R-CNN has been proposed to reduce the computational time of the R-CNN algorithm (Girshick, 2015). In this method, after taking an image as the input, a convolutional neural network should be applied to generate the ROI. Next, an ROI pooling layer is applied to all regions for reshaping them into a fixed size. Lastly, a softmax layer and linear regression should be used to output classes and bounding boxes, respectively.

3.4. Faster R-CNN

The major difference between Fast R-CNN and Faster R-CNN is that the former approach uses selective search to generate region proposals. In contrast, the latter leverages a Region Proposal Network (RPN) method, discussed extensively in Ren et al. (2015). Faster R-CNN takes an image as the input and passes it through the process, generating the feature map of the image. Then RPN is applied to these feature maps, returning the object proposals and objectness scores. An ROI pooling layer is applied to these proposals to bring down all the proposals to the same size. Finally, the proposals are passed to a fully connected layer with a softmax layer and a linear regression layer at its top, to classify and output the bounding boxes for objects. R-CNN-based algorithms require sub-regions and none of them consider a complete image to apply the process. Since these systems operate in consecutive stages, the performance of each stage heavily depends on the performance of its previous stage.

3.5. Mask R-CNN

Mask R-CNN algorithm is an extension of the Faster R-CNN, which adds a Mask network branch for ROIs prediction segmentation parallel to object classification and bounding-box regression (He et al., 2017). It involves two stages. The first stage consists of two networks including backbone and region proposal network. These networks run once per image to give a set of region proposals. In the second stage, the bounding boxes and object classes are predicted for each of the proposed regions obtained in the former stage.

3.6. YOLO

You Only Look Once, or YOLO, is one of the most popular object detection algorithms for real-time applications that often works best in terms of speed and outcome. It was first introduced by Redmon et al. (2016). Unlike region-based algorithms, this algorithm looks at the whole input frame. It then predicts how to identify, classify, and localize objects in the frame. This approach can be performed in real-time object detection since it is faster than region-based methods. YOLO takes an image as the input and divides it into S×S grids to predict if there is an object in that specific cell. Using this information, YOLO can predict the class of the objects. Unlike region-based methods, which require thousands of neural networks, in YOLO, the input frame passes through a single network evaluation. YOLO has three versions and in each new version, the creators made improvements in the detection accuracy compared to the previous one.

3.7. Single-shot multi-box detector (SSD)

SSD algorithm was first introduced by Liu et al. (2016). Unlike RPN-based approaches that needed two shots to generate region proposals, this method only needs one shot to detect objects. The SSD algorithm operates in two steps: extracting the feature maps and applying convolutional filters to detect objects. SSD is faster than region-based algorithms since it increases the speed by eliminating the need for region proposal networks. This elimination may affect the accuracy of the object detection, but SSD compensates for this drawback by applying improvements such as multi-scale features and default boxes and uses lower image resolution to achieve higher accuracy.

4. Review methodology

This review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Moher et al., 2010) to explore related literature in the field of deep learning-based object detection in augmented/mixed reality over the past ten years. Five databases were selected to search, including Scopus, Web of Science, IEEE Xplore, ACM, and ScienceDirect. These databases provided a wide variety of research studies in the corresponding fields.

4.1. Study selection

The search terms in this study included “deep learning” AND “object detection”, AND “augmented” OR “mixed reality”. These terms were used together to extract the studies that leveraged deep learning-based object detection in the context of augmented or mixed reality technology. The literature search was limited to the articles published in the past decade from 2011 to 2020. However, most of the articles directly related to this review were published after 2016. The initial search in this study resulted in 4835 papers, and after deleting duplicates, we ended up with 4136 unique articles. After screening the papers toward their title and skimming their abstract, 3727 papers were excluded from the study due to the irrelevancy of the titles and abstract. For the 409 remaining articles, we screened their full-text, and 69 research articles that met our inclusion criteria were selected as the most relevant papers for the current study's purpose and were included in the review. Fig. 1 shows the summary of the PRISMA process used for this study.

For this systematic review, the following information was extracted from the 69 papers:

- Types of deep learning algorithms
- Types of AR devices
- Computation platform
- Dependent and independent variables
- User study
- Year of publication
- Application or scope

5. Object detection in AR

Similar to many other areas, object detection has been extensively used in AR as well. Depending on the available amount of data, the configuration of the AR device, and the goal of detection, several types of algorithms can be used in AR applications. Traditional object detection in AR mainly consists of marker-based methods and statistical classifiers. One of the first studies on picture properties and pictorial pattern recognition was introduced by Rosenfeld and Pfaltz (1966), which elaborates on computer processing pictorial information techniques. Another early work in this field is an image processing approach proposed by decomposing an image into primitive pieces as a basis for reference component description (Fischler and Elschlager, 1973). These studies were primarily based on matching techniques and part-based algorithms. Later studies focused on classifiers such as Viola-Jones Haar Cascade, Histogram of Gradient (HOG), Scale-Invariant Feature Transform (SIFT), and Speeded-Up Robust Features (SURF). However, traditional object detection algorithms used in AR often suffer from issues such as lack of accuracy or being computationally intensive. Recent research in AR leveraged deep learning-based object detection to alleviate such challenges, either replacing traditional object detection methods or as a complementary component to mitigate their shortcomings.

5.1. Deep learning-based object detection in AR

Object detection based on deep neural networks has been extensively used in AR. Compared to the traditional techniques explained in Section 5, deep learning-based object detection is a novel and successful method. The main difference between the detection based on neural networks and statistical classifiers is that the user designs the latter. In contrast, in neural networks, the algorithm should learn a large sample of data to represent the response variable successfully. This section divides the previous works into several categories, including applications and computation platforms, and presents a comprehensive review of each category.

5.1.1. Applications

Previous studies on deep learning-based object detection in AR are centered on several areas including but not limited to education, manufacturing, aerospace, and robotics. Among all areas, manufacturing, autonomous vehicles, and assistance are more common.

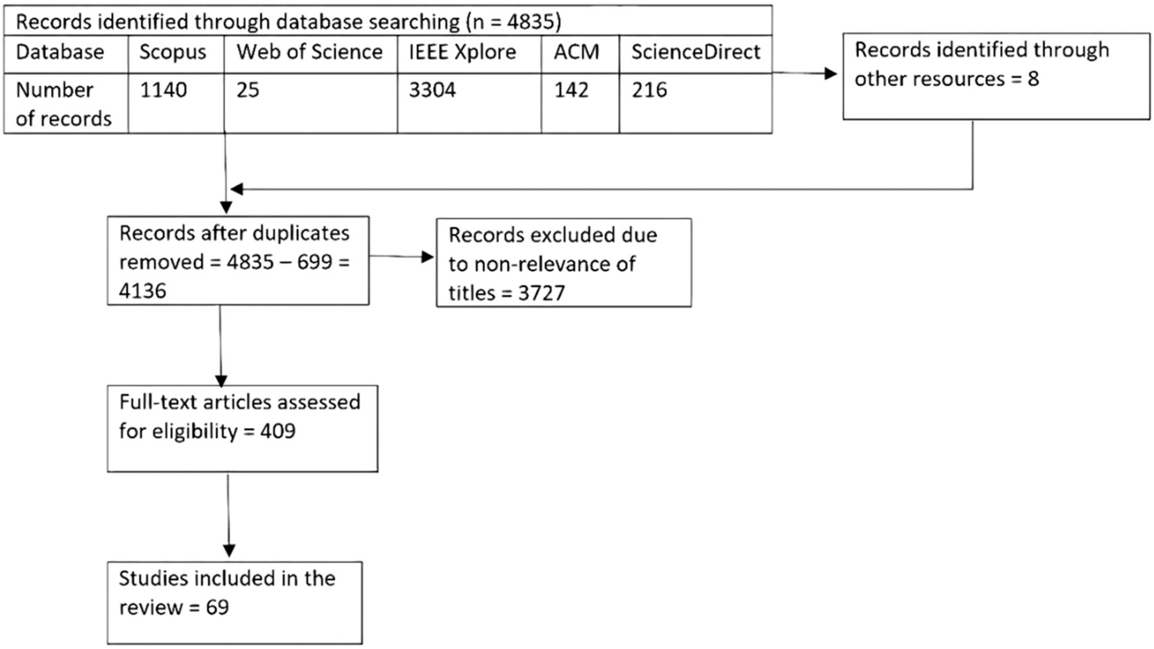

Records identified through database searching (n = 4835)
Database
Scopus
Web of Science
IEEE Xplore
ACM
ScienceDirect
Number of records
1140
25
3304
142
216
Records identified through other resources = 8
Records after duplicates removed = 4835 – 699 = 4136
Records excluded due to non-relevance of titles = 3727
Full-text articles assessed for eligibility = 409
Studies included in the review = 69

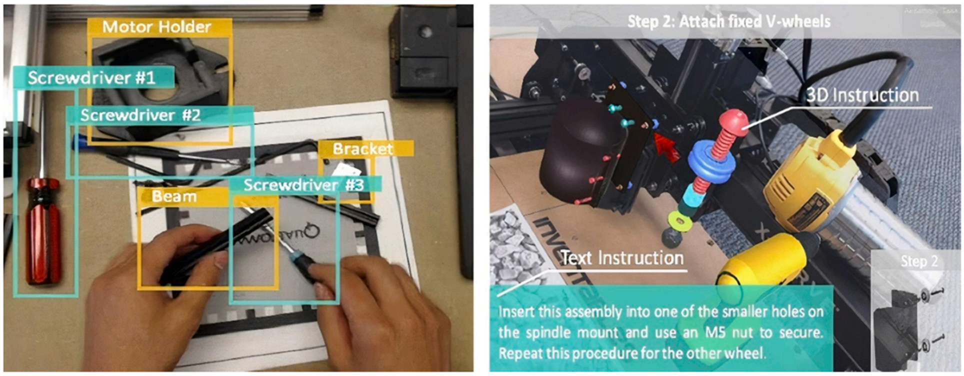
Fig. 1. PRISMA process for selecting related studies.

Also, some papers are not focused on specific areas and only provide a framework for object detection. Table 1 summarizes the applications used in the reviewed studies (Paper ID). Note that the Paper ID is used in the upcoming tables and figures to indicate each reviewed study. This section introduces the use of deep learning-based object detection in AR for the three most common applications including manufacturing, driving, and assistive technologies.

5.1.1.1. Manufacturing. A major application of deep learning-based object detection in AR is in the context of manufacturing. AR and computer vision have been used to make the factories smarter while facilitating the tasks for workers. Subakti and Jiang (2018) proposed a mobile AR that recognizes three different industrial machines and their components. Digital information is sent to the smartphone to be superimposed on the machine images shown on the display. By leveraging the touch screen and distance perception of smartphones, this system can provide two modes of interaction, including touch and distance-aware interactions. Assembly is a major task being carried out in manufacturing environments. To reduce the issues of manual assembly, S. Wang et al. (2018) proposed a new assembly

Table 1
Application areas used in the reviewed studies.

Applications	Study [Paper ID]
Manufacturing	Ramakrishna et al. (2016) [3] Subakti and Jiang (2018) [11], S. Wang et al. (2018) [12], Židek et al. (2019a) [20], Židek et al. (2019b) [24], Tao et al. (2019) [27], Corneli et al. (2019) [28], Sun et al. (2019) [36], Kim and Lee (2019) [41], Zheng et al. (2020) [42], Park et al. (2020a) [46], Park et al. (2020b) [47], Lai et al. (2020) [48], Kästner et al. (2021) [50], Konstantinidis et al. (2020) [55]
Navigation assistance for elderly, disabled people, and shopping	Advani et al. (2017) [2], Eckert et al. (2018) [10], Lin et al. (2018) [15], Cruz et al. (2019) [22], Park et al. (2019) [25], Fuchs et al. (2019) [34], McKelvey et al. (2019) [35], Fuchs et al. (2020) [57], Nilwong and Capi (2020) [61]
Driving	Abdi and Meddeb (2017) [8], Abdi et al. (2017) [9], Abdi and Meddeb (2018) [13], Alhaija et al. (2018) [14], Anderson et al. (2019) [32], Pai et al. (2020) [43], Deore et al. (2020) [49], Zhou et al. (2020) [54]
Robotics	R. Wang et al. (2018) [16], De Gregorio et al. (2019) [26], Farasin et al. (2020) [44], Kästner et al. (2020) [45]
Education	Karambakhsh et al. (2019) [21], Huynh et al. (2019) [37], Plecher et al. (2020) [59]
Healthcare	Wang et al. (2019) [38], Waithe et al. (2020) [67]
Search and rescue	Llasag et al. (2019) [39]
Firefighting	Bhattarai et al. (2020) [53]
Geometric perception	Han et al. (2020) [62]
Sudoku puzzle	Syed et al. (2020) [66]
Children guessing game	Putze et al. (2020) [69]
Geovisualization	Rao et al. (2017) [4]
Obstacle detection	Połap et al. (2017) [6]
Virtual agent positioning	Lang et al. (2019) [31]
Not specified	Tobías et al. (2016) [1], Ran et al. (2017) [5], Sutanto et al. (2017) [7], Liu and Han (2018) [17], Mahurkar (2018) [18], Rodrigues et al. (2018) [19], Bahri et al. (2019) [23], Liu et al. (2019) [29], Apicharttrisorn et al. (2019) [30], Huang et al. (2019) [33], Li et al. (2019) [40], X. Li et al. (2020) [51], Hu et al. (2020) [52], Rathnayake et al. (2020) [56], Le et al. (2020) [58], Ahn et al. (2020) [60], Dasgupta et al. (2020) [63], Golnari et al. (2020) [64], Cheng et al. (2020) [65], Lomaliza and Park (2020) [68]



Insert this assembly into one of the smaller holes on the spindle mount and use an M5 nut to secure. Repeat this procedure for the other wheel.

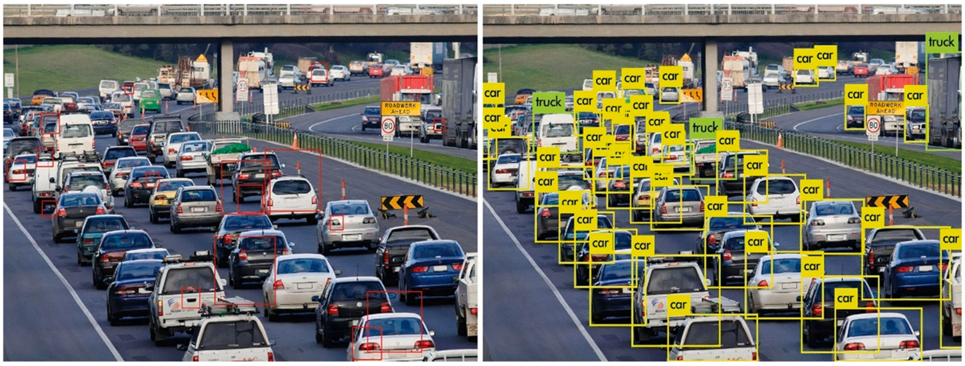
(a) Assembly part detection

(b) Instructions in AR display

Fig. 2. Results of tool/part detection and providing instructions in manufacturing using R-CNN algorithm (Tao et al., 2019).

fault detection based on deep learning and mixed reality (MR), which requires training a pretreatment model and detecting targets via deep learning and extracting feature information. This method could significantly improve equipment efficiency and reduce assembly errors. Židek et al. (2019b) used deep learning to identify assembly parts and speed up the assembly process with AR application and dynamic recognition, demonstrating the potential of their approach for improving the assembly tasks. They also presented a methodology for speeding up the CNN training process based on the automated generation of input sample data for learning without any monotonous manual work. This way, it would significantly shorten sample preparation time without automation. A portable visual device based on binocular vision and deep learning was developed by Zheng et al. (2020) to realize fast detection and recognition of cable brackets that were installed on aircraft airframes. It consisted of three subsystems: bracket inspection, cable text reading, and assembly process guidance based on AR. It could assist workers in quickly inspecting the state of brackets by showing the installation path of cables to be assembled. This approach has improved the assembly efficiency and quality of the aircraft cable assembly process. A worker-centered training and assistant system for intelligent manufacturing was proposed in Tao et al. (2019). The worker's state was perceived with multi-modal sensing and deep learning methods, and was used to determine the potential guiding demands. Then, active instructions with AR were provided to suit the worker's needs. The experiment showed the feasibility and promising results of applying the proposed system for training and assisting frontline workers. Park et al. (2020a) introduced a deep learning-based mobile AR for smart task assistance and in a further study, they proposed a user-centered AR method that proved to be faster than the marker-based AR while overcoming the limitations of existing interactions in wearable AR such that complex tasks can be performed more accurately and effectively (Park et al., 2020b). An AR instructional system integrated with Faster R-CNN for the mechanical assembly was proposed in Lai et al. (2020). A synthetic tool dataset was developed using data augmentation with CAD models and successfully deployed to detect real tools. The experimental results on the assembly task indicated a considerable improvement in the assembly performance, compared to the conventional methods. An AR-based human assistance system for complex manual tasks incorporating deep neural networks was proposed by Kästner et al. (2020). AR was combined with object and action detectors in this study to make workflows more intuitive and flexible. In another study, AR and computer vision (CV) techniques were utilized to support novice operators in the maintenance procedures. A mobile AR maintenance assistant using a handheld device's camera was introduced by Konstantinidis et al. (2020) to generate AR maintenance instructions. The performance of this system showed promising results in a real-world scenario. Fig. 2 provides an example of (a) assembly part detection using AR and deep learning as well as (b) task instructions in AR display.

5.1.1.2. Driving and autonomous vehicles. Deep learning-based object detection has been used in driving and autonomous vehicles for obstacle avoidance and increasing drivers' performance by enhancing their awareness of the environment. AR HUD and deep learning can be used to recognize road obstacles and interpret and predict complex traffic situations, which can significantly improve the driving experience (Abdi and Meddeb, 2017). A real-time approach for traffic sign recognition has been employed using deep learning. This approach improves the accuracy of the traffic sign detector to assist the driver in various driving situations, increase driving comfort, and reduce traffic accident risks. Experimental results showed that the suggested method was comparable to state-of-the-art approaches but with less computational complexity and shorter training time. It was also mentioned that AR impacts the allocation of visual attention more strongly during the decision-making phase (Abdi and Meddeb, 2018). In a study by Alhaija et al. (2018), synthetic data in urban driving scenes were generated by combining AR and computer vision suitable for training deep neural networks. However, synthetic objects can only be placed on top of real images. They thus cannot be partially occluded by the real objects. In addition, Deore et al. (2020) developed an algorithm to combine deep learning and AR in the context of autonomous vehicles. The trained deep learning test model performed well on detecting the AR artificial navigational signs. As the objects were clear compared to real signs it was easier for the algorithm to detect. However, the artificial objects created using AR during testing hid important surrounding details. The detail in real self-driving vehicles can be any human or moving object which should be detected to avoid an accident. An augmented reality environment for drivers where important information is displayed in Holograms was proposed in Anderson et al. (2019), where real-time object and lane detection was studied to enhance the driver's ability to avoid collisions. A driver assistance system that uses a network model based on deep learning technology was developed by Pai et al. (2020). A camera was used



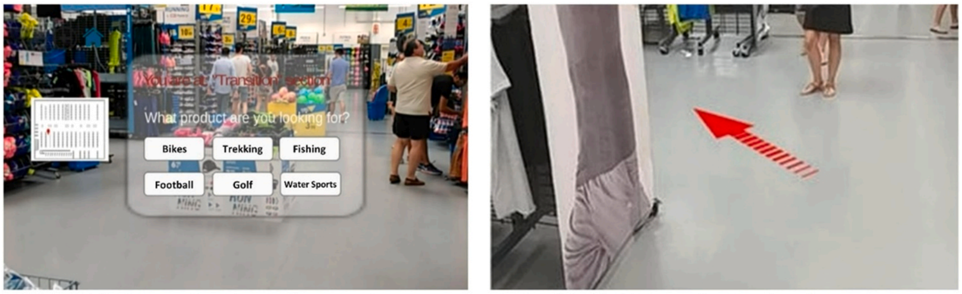
(a) Object detection results using Viola-Jones (b) Object detection results using YOLO

Fig. 3. Results of traditional and deep learning-based object detection algorithms used in the autonomous driving study (Anderson et al., 2019).

to capture the vehicle's image ahead and was used to predict the type and position of the forward object and the ground road signs using trained network models. This gave the driver a view of the road ahead with information about possible hazards and warnings of danger to enhance safety. Fig. 3 shows an example of object detection results using (a) a traditional method (Viola-Jones) and (b) a deep learning-based method (YOLO).

5.1.1.3. Assistance. AR and its applications with deep learning were not only used in workplaces or modern technologies, but they have also been used to assist people in their daily routine activities. Some critical applications of these assistive approaches pertain to helping the elderly or enhancing navigation for visually impaired people. Using deep learning approaches including pose estimation, object and face detection, and a spatial AR technology, Park et al. (2019) provided alerts and assistance for daily works of older adults in their real environment. In addition, deep learning-based object detection in AR can be leveraged to enhance the navigation experience for all people regardless of their impairment, especially when navigating a new environment for the first time. A novel campus navigation app that uses AR to provide users with a new way was introduced in Lin et al. (2018). Using the combination of AR and deep learning-based object recognition, the information about the campus environment was overlaid in the real world, making an interactive interface. To improve the app efficiency, this paper presented a virtual terrain modeling interface with deep learning to improve the object recognition ability. In addition, studies were conducted to address robot navigation as well. Nilwong and Capi (2020) presented a deep reinforcement learning-based robot outdoor navigation method using visual information. Experimental results from the simulation showed the high effectiveness of the navigation system inside the simulation. The real experiments showed a potential of the game-based Deep-Q Network (DQN) and a simple marker-based AR method for simple navigation tasks in short distances. The implemented DQN was trained in a game-based simulation environment, then directly employed to the real robot without any changes.

To enhance the shopping experience and guide visually impaired people, Advani et al., (2017) developed a system to provide tactile feedback from a custom glove equipped with a camera and vibration motors as well as auditory and visual feedback from a pair of smart glasses. They described the various features incorporated into this visual-assistance system in multiple contexts while highlighting the efficiency of personal visual assistance systems in day-to-day activities. A system showing AR models to the users directly at the store was proposed by Cruz et al. (2019). This system provides user navigation and improves the localization of certain products at the store. However, some limitations such as requiring an Internet connection, high power consumption on mobile devices, dynamically changing environment, and localization difference between the taken picture and system response may cause failures to guide the user accurately. MR headset-mediated technologies on food and grocery selection are feasible, but little is known about their impact on user choice and other outcomes in the real world and users' opinions on the efficiency of such systems. Fuchs et al. (2020) presented a novel framework that combines research streams into a novel user support system for providing healthy food choices to address the research gap of the joint application of CV-based


You are at "Transition" section
What product are you looking for?
Bikes
Trekking
Fishing
Football
Golf
Water Sports

(a) Detection in real scenario (b) Guidance for navigation

Fig. 4. An example of object detection to assist with shopping navigation using ResNet (Cruz et al., 2019).

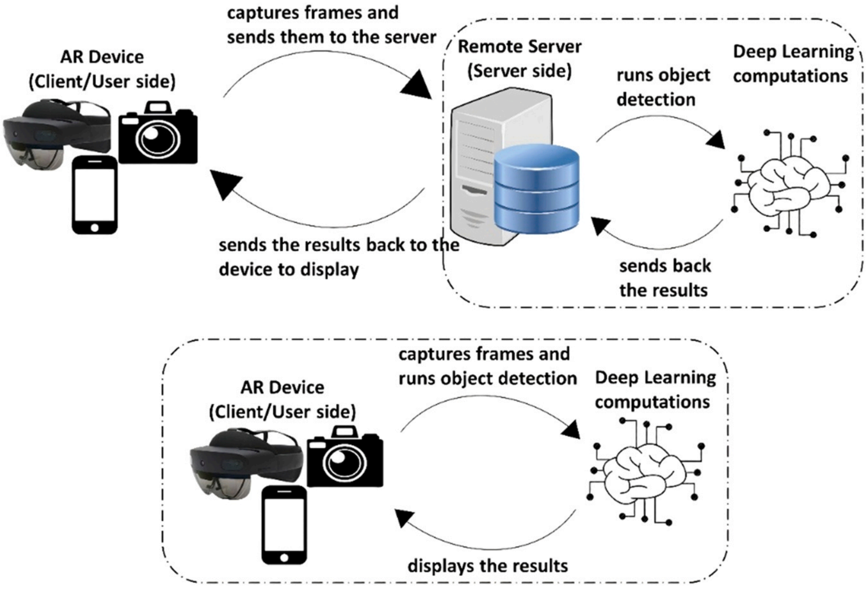

captures frames and sends them to the server
AR Device (Client/User side)
Remote Server (Server side)
runs object detection
Deep Learning computations
sends the results back to the device to display
sends back the results
AR Device (Client/User side)
captures frames and runs object detection
Deep Learning computations
displays the results

Fig. 5. Two processes of the object detection computation: (a) remote server-based process, (b) local device-based process.

detection of food items. They used an MR wearable headset to explore the technical feasibility and potential impact of MR food labels in affecting beverage and food purchasing choices. They assessed whether visual cues in the form of front-of-package labels (i.e., Nutri-score) influence consumers in preferring and selecting healthy or unhealthy beverages and foods and analyzed consumers with low food literacy. They also included an in-depth discussion on the latency of product detection via CV to assess the technical feasibility of detecting packaged products under realistic circumstances. Fig. 4 shows an implementation of AR and deep learning for enhancing (a) shopping experience and (b) navigation in large retail stores.

5.1.2. Computation platform (local vs. server)

Deep learning is capable of making the AR/MR systems smarter. However, to implement deep learning-based object detection, some obvious considerations need to be considered when choosing between a remote server or a local device. Fig. 5 shows the two processes for the object detection computation. This section discusses some advantages and disadvantages of each approach and explains some use cases from the literature. Capability, computation cost, complexity, and size of the model are important factors in choosing the computation method. Table 2 provides a summary of computation platforms used in each reviewed paper.

For the remote server-based computation, the client/user (i.e., AR device) captures frames and sends them to the remote server; then the model processes the data on the server. Then, the model sends the output back to the client/user. Network connection and delays are important factors during this process. Implementing operation in an environment with high delay can increase the object detection latency. However, this method is generally faster than using a local device since it can process a higher amount of data in a shorter time. Due to the high memory bandwidth and the ability to conduct numerous parallel computations, GPUs have become a widely accepted method for training deep learning models. The quality of the GPU can highly affect the performance of deep learning models. In contrast, using a local device to implement deep learning computations is more convenient and flexible, requiring lighter algorithms to be executed effectively. Otherwise, time and accuracy should be sacrificed.

Table 2
Computation platforms used in the reviewed studies.

	Server	Local	Server and local (both)
Paper ID	2, 3, 6, 7, 10, 11, 12, 17, 18, 19, 21, 22, 23, 27, 29, 30, 34, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 60, 62, 63, 64, 65, 66, 67, 68	8, 9, 13, 14, 15, 16, 20, 24, 25, 26, 28, 31, 32, 35, 59, 61, 69	1, 4, 5, 33

Out of 69 studies collected for this review, 48 studies used cloud, edge, or GPU servers. 17 studies implemented their computation on the local AR devices, and 4 studies used both server and local devices for their computations. Different methods and approaches were used to implement object detection either on a server or on a local device.

Since mobile devices have become relatively powerful to handle the computations near real-time, Tobías et al. (2016) proposed a method in which three lightweight CNN algorithms including AlexNet, GoogLeNet, and NIN are used for object recognition on GPU, CPU, and mobile devices. AlexNet requires large memory storage, making it less desirable for implementation on mobile devices. In contrast, GoogLeNet and NIN models require lower memory. This study showed that GoogLeNet led to the longest processing time while Network In Network (NIN) model demonstrated the least processing time. A near real-time mobile outdoor AR was proposed by Rao et al. (2017) . They used the SSD algorithm, a vision-based approach that can use natural features of geographic objects under various conditions. However, using this method in outdoor environment under poor signal conditions is challenging. Since SSD is still too slow to efficiently perform the required computations without a powerful GPU, they modified this algorithm to a lightweight SSD to make it more mobile-friendly by providing a less accurate base network, fewer feature layers, and smaller input sizes. Compared to the original SSD and a fast YOLO approach, this method was more robust while having less mean average precision (mAP).

A new approach for object detection using deep learning networks trained remotely by 3D virtual models was proposed in Židek et al. (2019a), where Faster R-CNN Inception v2 was selected due to its high accuracy. However, considering the disadvantage of its network size and recognition speed, it was not suitable for embedded devices. MobileNet v2 reduced CNN, was more appropriate for embedded devices, since it reduces the processing delay and is optimized for low-performance computing devices such as smart glasses with Android OS. This study showed that AR devices with embedded processing units could reach a decent amount of frames per second which is satisfactory if the task does not require too many movements. Other variations of deep learning algorithms can be used to comply with less powerful devices processors. In Corneli et al. (2019), a lightweight variation of YOLO (i.e., tiny YOLO v2) was used to perform the whole process on site and in real-time. Bhattarai et al. (2020) proposed an embedded system where the processed images were analyzed and returned through wireless streaming in real-time using an embedded GPU development platform. A quantized version of the SSD MobileNet v2 has been used because this network had a suitable speed of detection for AR applications, and it hadana mAP ideal for the dataset they used. A quantized network means that smaller memory space is required for the weights and the loading times of model are faster, which is necessary for mobile applications (Plecher et al., 2020). In a study by Sutanto et al. (2017) they proposed a markerless AR using a Faster R-CNN algorithm and a mobile device capturing the images of objects and sending them to the server where the computations are performed. The results were displayed on the mobile, based on the stored object detection in the database and they could be seen through a 3D lenslet array case. By incorporating 3D integral imaging, they successfully implemented this application on non-high-specs devices. Lang et al. (2019) used HoloLens to perform the user's computations. However, due to the limited computing power of HoloLens, they ran optimizations on a PC using other processors.

Performance is an essential factor when an object detection model is implemented. A single network evaluation CNN predicted region of interest and class probabilities directly from full images in one evaluation (Eckert et al., 2018). This method significantly improved performance over the other state-of-the-art models. An android application to enable real-time AR was developed to perform object detection (Ran et al., 2017). They performed the object detection on both server and mobile devices and showed that factors affecting detection performance in these scenarios include model size, offloading decision, and video resolution. While the results using the server highly depend on the network condition, they concluded that offloading on the server improved frame rate and accuracy. Since time consumption is considered one of the main challenges of deep learning, studies used servers to increase the speed of computations even in the training process (Karambakhsh et al., 2019). Advani et al., (2017) used a high-performance cloud computing technique to leverage both GPU and field-programmable gate arrays (FPGA). While they could accelerate the process by exploiting parallel algorithms, the server could not meet the real-time computations since it should have handled multiple connections at once. To achieve real-time processing, the system must leverage all computing powers available at edge devices and local infrastructure. A server-based object detection has been implemented using near real-time deep learning. A pre-trained model of YOLO v2 was used to perform the object detection on the server. Liu et al. (2018) showed the importance of characterizing tradeoffs between augmentation quality and latency when implementing on the edge server. They also implemented a protocol to maximize augmentation quality under varying network conditions and computation workloads. Bahri et al. (2019) developed an object detection system to recognize the objects via HoloLens and applied the YOLO algorithm at the server side to transmit the data from the user or client sides. To increase the detection speed on the server side and display results to the user, they created an algorithm between HoloLens as a client and a desktop as a server using Transmission Control Protocol/Internet Protocol (TCP/IP). Huang et al. (2019) proposed a framework using both server and local methods. High computation complexity tasks were offloaded to the edge servers and low complexity tasks were executed on mobile devices or the edge server depending on the network latency. However, dynamic network condition changes made the process unstable and the limitations of the on-device deep learning models were considerable. To control these factors, a cache and matching algorithm was designed on the mobile device to enhance the performance of the recognition tasks. This solution improved the quality of the mobile AR application.

Table 3
Input data used in the reviewed studies.

	Images/Videos	Point clouds
Paper ID	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 63, 64, 65, 66, 67, 68, 69	45, 62

Based on the reviewed studies and the tradeoffs each method entails, server-based object detection is more common and easier to implement. However, there are still some drawbacks that should be considered. Although the difficulties during the implementation process may differ case-by-case, the limitations and capabilities should be identified and tested before implementation.

Cost and latency due to network conditions are two critical challenges that one may face. Predicting and providing real-time object detection, especially on relatively more complex video frames, is more challenging, and there is no guarantee that the results will be received in real-time. Therefore, it can negatively affect the user experience in those situations. On the other hand, client-side models are much cheaper than server-side models and in theory, they should have lower latencies since they do not send and receive requests to a server. All computations will be performed in one place, which is the client-or user-side. However, in practice, due to the hardware limitations of existing devices, the latencies can actually be more significant than those of the server-side. They could be more challenging to implement since many optimization operations should be performed to allow the system to run smoothly without any hindrance.

6. Summary of review results

For this study, publications of the last ten years from 2011 to 2020 were extracted from different resources. The papers were introduced regarding the type of input used for object detection, evaluation metrics, publication year, type of AR devices, and type of algorithms. Table 3 shows the type of inputs used in reviewed studies. The papers mainly employed frames of videos or images where the content should be augmented. In these cases, the AR device records and sends the videos of the environment to the object detection algorithm in real-time. The object detection and augmentation will be performed based on the frames captured from the streaming videos. Another less frequently used input includes point clouds or 3D object detection instead of 2D images. This method provides an efficient way for localizing, characterizing, and obtaining depth information, such as measuring distance using 3D data of the objects. After registering all of the detected point clouds on the scene, complete capture of the scene can be obtained. As

Table 4
Metrics used in the reviewed studies.

Evaluation type	Metrics
Computation efficiency	**Runtime** [1, 4, 9, 10, 15, 18, 19, 23, 17, 32, 33, 36, 39, 41, 44, 52, 57, 62, 63, 67, 68]
	Latency [4, 5, 9, 13, 24, 29, 33, 52, 54, 55, 56, 60]
	Energy consumption [1, 2, 5, 29, 30, 54, 56, 60]
Performance	**Detection accuracy** [1, 2, 3, 4, 6, 7, 8, 10, 11, 12, 15, 21, 22, 25, 27, 29, 30, 34, 35, 36, 40, 41, 42, 47, 49, 51, 52, 53, 57, 60, 61, 64, 66, 68, 69]
	Precision [4, 5, 8, 9, 13, 14, 17, 20, 23, 24, 25, 26, 28, 32, 33, 39, 42, 43, 44, 48, 51, 55, 56, 57, 58, 59, 67]
	Recall [6, 8, 9, 12, 26, 32, 42, 43, 49, 51, 56]
	Error rate [6, 16, 22, 29, 37, 38, 46, 48, 50, 61, 65, 68]
	Loss [4, 12, 18, 31, 45, 49, 60, 64]
	Intersection Over Union (IOU) [26, 30, 62, 66]
	Task completion time [3, 27, 47, 48, 50],
	Users' accuracy [27, 48]
	Users' error [27, 50, 69]
Subjective measurements	31, 46, 47, 50, 57, 69

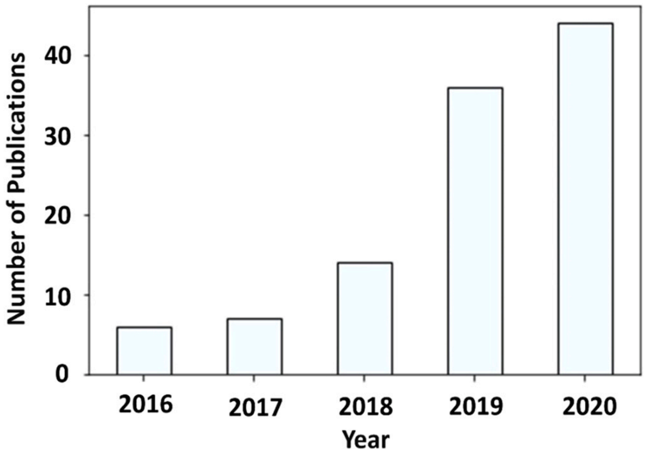


Fig. 6. Number of publications per year.

shown in Table 3, image/video inputs are a more common input than point clouds since they are easier to capture and analyze.

For evaluating the proposed object detection methods in AR, reviewed studies generally focused on three evaluation approaches including computation efficiency measures such as runtime and latency, performance measures such as accuracy and error for both systems and users, as well as subjective measures such as user acceptance, cognitive workload, ease of use, enjoyment, and usefulness using self-reported surveys only for studies that evaluated their proposed approaches with human subjects. Table 4 summarizes evaluation metrics used in the reviewed studies. Each evaluation method has its advantages and disadvantages. Using a combination of efficiency, performance, and subjective measurements provides a more robust and comprehensive evaluation.

Fig. 6 shows the number of studies published each year. The results show that deep learning-based object detection in AR was not studied before 2016. Since 2016, this topic has received remarkable attention and the number of publications has been increasing over the years. It can be observed that the highest number of published papers in this field corresponds to the year 2020. It should be noted that the concept of deep learning-based object detection was introduced in 2014, and before that, traditional methods, such as Viola-Jones, HOG, and SVM were used. The current trend of this topic shows the interest of researchers in solving associated problems using these technologies.

Across the 69 studies reviewed in this paper, six different categories of AR devices including five distinct categories for displaying AR information, were identified: Wearable devices, Projection-based AR (including HUDs), Monitors, and RGB Cameras. The "Other" category includes devices that are only used for specific applications. Wearable devices such as HoloLens are the most frequently used devices while mobile devices come second. These two categories were used in more than half of the papers. It could result from the quality and simplicity of these devices for providing interactive experiences and the availability of mobile devices for AR applications. Fig. 7 shows the summary of devices used in studies per year. It can be observed that the use of AR devices is increasing especially for wearables and mobile devices, while projection-based AR shows a cyclic pattern. Moreover, use of camera and monitor is not showing a specific pattern each year, but it can be seen that their use has also been increasing over the years.

The combination of deep learning algorithms and devices used in the 69 collected papers is demonstrated in Fig. 8. Based on this result, among all devices and algorithms, YOLO has been used more

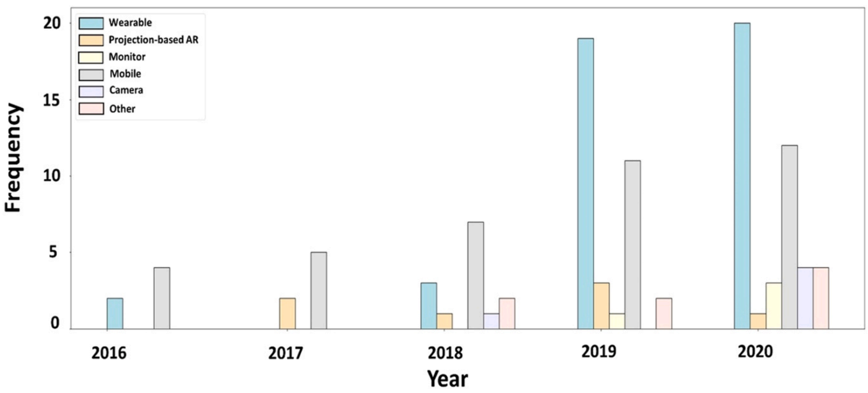


Fig. 7. Number and type of devices used per year.

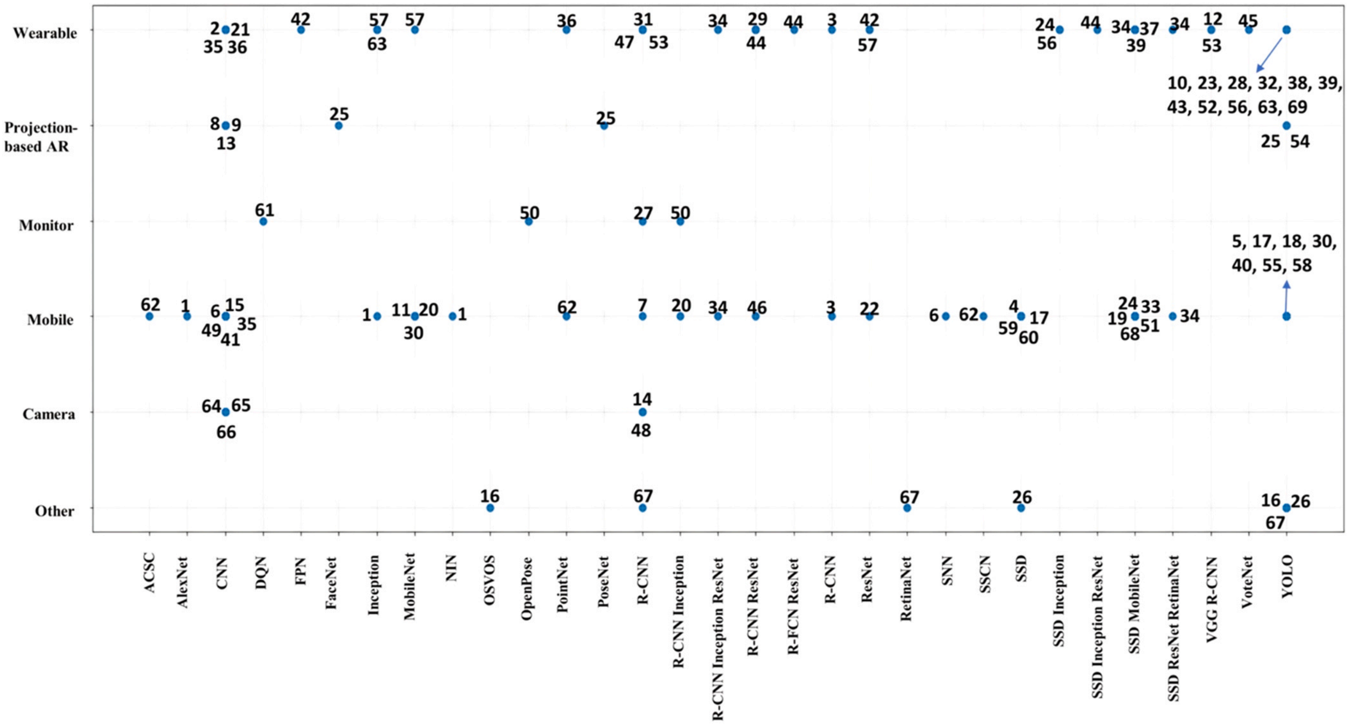


Fig. 8. Combination of algorithms and devices used in reviewed papers.

Table 5
Studies conducted human subject experiments.

Paper ID	The number of human subjects
3, 27, 48, 56, 69	20
31, 50	30
44	10
46	20, 20 (two separate experiments)
47	25, 20 (two separate experiments)
57	61

than other deep learning algorithms, having a dense intersection with both wearables and mobile devices.

Out of 69 papers, only 11 papers mentioned that they conducted user studies and evaluated their proposed systems with human subjects (Farasin et al., 2020; Fuchs et al., 2020; Lai et al., 2020; Park et al., 2020a, 2020b; Tao et al., 2019; Kästner et al., 2020; Lang et al., 2019; Putze et al., 2020; Ramakrishna et al., 2016; Rathnayake et al., 2020). As shown in Table 5, the studies involve 10–61 participants. However, in all of the conducted studies, the number of female participants were significantly less than the number of male participants, which can possibly make the results biased toward one gender, especially in the sense that the research showed that females and males have different reactions when using the AR technology (Dirin et al., 2019). Surprisingly, most of the studies did not mention participants' age ranges, but from a few existing data, their ages ranged from 18 to 50 with an average of 25. Also, most of the studies did not clarify whether the user population was representative of the real target users. However, based on the available information from some studies, we can conclude that most of them recruited university students who were not necessarily representative of a larger population of users.

7. Discussion

Based on the existing data in the field of deep learning-based object detection in AR, it can be observed that there are still many challenges that should be addressed to enhance the performance of the detection algorithms in the future. This section provides an overview of the challenges we observed based on this review.

A powerful computation resource, a large training dataset, and a suitable machine learning model can significantly improve deep learning performance. However, implementation issues on mobile devices remains a challenge that need to be optimized to reduce computational time and improve the effectiveness of the algorithms to make them more lightweight, fast, and accurate. In addition, mobile applications generate noisy data if the algorithm is not robust enough and the dataset is small. In general, energy consumption is the primary concern for mobile AR applications. Additional work must be done to reduce the detection and segmentation execution time. In addition, several datasets can be used as a benchmark for object detection to train algorithms. However, many studies still use manual labeling that is not cost- and labor-effective. Another challenge is dealing with the limitations of wearable devices such as HoloLens. HoloLens's hardware is unpleasant and uncomfortable for prolonged durations of wear. Also, HoloLenshas a limited battery capacity and sometimes requires continuous and stable access to the network. These limitations may hinder its performance in many real-world or in-the-wild applications. In addition, due to the limited range of the depth sensor of HoloLens, when identifying an object at a long distance, it is necessary to move closer to the target to create a 3D mesh of the physical space before detection. Another limitation that needs more improvement is detection accuracy when there is a reflective medium or a shiny surface such as glass in the environment. The detection also may fail when the object being scanned moves.

The most important limitation of previous studies is the lack of human subject experiments. Many of the earlier efforts did not evaluate the performance of the proposed systems in AR using human subjects. Park et al. (2019) mentioned the importance of usability and user experience when designing a system for humans and considered testing their system with users in future studies. Similarly, as a future direction for a driving-related task aiming to increase drivers' awareness, Anderson et al. (2019) discussed the necessity of conducting a user study to evaluate the impact of holograms on drivers. Moreover, Park et al. (2021) proposed the limitations and future directions of their research aiming to use subjective approaches to evaluate the physical and mental workload of the users in a human-robot collaboration scenario. Hu et al. (2020) also noted the importance of the interaction modes in their proposed system. To provide a better user experience in terms of interaction modes, further user studies need to be conducted to evaluate the system. Running experiments with human subjects could help evaluate the systems from humans' perspectives; after all, all these applications are ultimately developed to be used by humans. Conducting user studies can also reveal whether the system is suitable for users to use in their everyday lives. Evaluating a system with human subjects can also lead to identifying the strengths and weaknesses of proposed systems. Most of the previous studies either did not include such evaluations in their research or conducted their evaluations with a small number of participants, making it difficult to validate the results. In addition, researchers should be more attentive to balancing the number of male and female participants when conducting a user study. In most existing research studies, the number of male participants is significantly greater than females. Moreover, while some research works conducted field studies including in-the-wild or in-situ, most of the studies performed their task in controlled lab environments with consistent lighting, static workplaces, and a limited number of objects which may not reflect the true utility and effectiveness of their approaches. Since deep learning-based object detection in AR is applied to many real-world situations, it is worth exploring them in a more uncontrolled format to better understand their usefulness in real tasks and environments.

8. Conclusion

This paper provided a comprehensive review of deep learning-based object detection in AR, including an overview of current technologies and devices in AR as well as frequently used algorithms for object detection. This review showed how deep learning is different from statistical classifiers for object detection and provided many advantages of using deep learning over traditional detection methods. It also represented the current state of deep learning-based object detection in AR regarding different applications and implementation methods. It was observed that the number of publications in this field is increasing, making this area a pervasive field of study. Depending on the type of algorithm, model size, network conditions, and the computing power of AR devices, care must be taken when implementing computations on the server-side or on the local devices Future studies in this field can be improved by designing more powerful mobile devices that can process the computations locally in real-time and designing more pleasant wearable devices for long durations of use. There is room for developing more lightweight devices in the future. In addition, depending on the application, methods such as inertial odometry for estimating the pose and optical flow can help to reduce energy consumption and processing time, respectively. Whether using a remote server or a local device, the tradeoffs between computation time, accuracy, latency, and battery drain should be made according to the task and AR device characteristics. To deal with manual labeling, increasing the number of datasets with reliable and accurate labels that use fewer samples for the learning stage can be a viable solution to enhance the accuracy and speed of detection. In general, for future studies, after taking into account the considerations mentioned above, it is worth it for researchers to investigate their proposed approaches in uncontrolled or wild environments while testing them on human subjects to validate their usability for real users and applications.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the Republic of Korea's MSIT (Ministry of Science and ICT), under the High-Potential Individuals Global Training Program (2020-0-01532 & 2021-0-01548) supervised by the IITP (Institute of Information and Communications Technology Planning & Evaluation).

References

Abdi, L., Meddeb, A. , 2017, April. Deep learning traffic sign detection, recognition and augmentation. In: Proceedings of the Symposium on Applied Computing. pp. 131–136.

Abdi, L., Meddeb, A., 2018. Driver information system: a combination of augmented reality, deep learning and vehicular Ad-hoc networks. Multimed. Tools Appl. 77 (12), 14673–14703.

Abdi, L., Takrouni, W., Meddeb, A. , 2017, June. In-vehicle cooperative driver information systems. In: Proceedings of the 2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC). pp. 396–401. IEEE.

Advani, S., Zientara, P., Shukla, N., Okafor, I., Irick, K., Sampson, J., Datta, S., Narayanan, V., 2017. A multitask grocery assist system for the visually impaired: smart glasses, gloves, and shopping carts provide auditory and tactile feedback. IEEE Consum. Electron. Mag. 6 (1), 73–81.

Ahn, J., Lee, J., Niyato, D., Park, H.S., 2020. Novel QoS-guaranteed orchestration scheme for energyefficient mobile augmented reality applications in multi-access edge computing. IEEE Trans. Veh. Technol. 69 (11), 13631–13645.

Alhaija, H.A., Mustikovela, S.K., Mescheder, L., Geiger, A., Rother, C., 2018. Augmented reality meets computer vision: efficient data generation for urban driving scenes. Int. J. Comput. Vis. 126 (9), 961–972.

Anderson, R., Toledo, J., ElAarag, H. , 2019, April. Feasibility study on the utilization of Microsoft HoloLens to increase driving conditions awareness. In: Proceedings of the 2019 SoutheastCon. pp. 1–8. IEEE.

Apicharttrisorn, K., Ran, X., Chen, J., Krishnamurthy, S.V., Roy-Chowdhury, A.K. , 2019, November. Frugal following: Power thrifty object detection and tracking for mobile augmented reality. In: Proceedings of the 17th Conference on Embedded Networked Sensor Systems. pp. 96–109.

Bahri, H., Krčmařík, D., Kočí, J. , 2019, December. Accurate object detection system on hololens using yolo algorithm. In: Proceedings of the 2019 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO). pp. 219–224. IEEE.

Bhattarai, M., Jensen-Curtis, A.R., Martínez-Ramón, M. , 2020, December. An embedded deep learning system for augmented reality in firefighting applications. In: Proceedings of the 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 1224–1230. IEEE.

Chen, I.Y., MacDonald, B., Wünsche, B. , 2008, December. Markerless augmented reality for robots in unprepared environments. In: Proceedings of the Australasian Conference on Robotics and Automation. ACRA08. pp. 3–5.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L. , 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062.

Cheng, Q., Zhang, S., Bo, S., Chen, D., Zhang, H., 2020. Augmented reality dynamic image recognition technology based on deep learning algorithm. IEEE Access 8, 137370–137384.

Choi, S.H., Park, K.B., Roh, D.H., Lee, J.Y., Mohammed, M., Ghasemi, Y., Jeong, H., 2022. An integrated mixed reality system for safety-aware human-robot collaboration using deep learning and digital twin generation. Robot. Comput. -Integr. Manuf. 73, 102258.

Chowdhury, S.A., Arshad, H., Parhizkar, B., Obeidy, W.K., 2013. Handheld augmented reality interaction technique. Proceedings of the International Visual Informatics Conference. Springer, Cham, pp. 418–426.

Corneli, A., Naticchia, B., Carbonari, A., Bosché, F., 2019. Augmented reality and deep learning towards the management of secondary building assets. ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction, vol. 36. IAARC Publications, pp. 332–339.

Cruz, E., Orts-Escolano, S., Gomez-Donoso, F., Rizo, C., Rangel, J.C., Mora, H., Cazorla, M., 2019. An augmented reality application for improving shopping experience in large retail stores. Virtual Real. 23 (3), 281291.

Dasgupta, A., Manuel, M., Mansur, R.S., Nowak, N., Gračanin, D. , 2020, March. Towards real time object recognition for context awareness in mixed reality: a machine learning approach. In: Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). pp. 262–268. IEEE.

De Gregorio, D., Tonioni, A., Palli, G., Di Stefano, L., 2019. Semiautomatic labeling for deep learning in robotics. IEEE Trans. Autom. Sci. Eng. 17 (2), 611–620.

del Amo, I.F., Erkoyuncu, J.A., Roy, R., Palmarini, R., Onoufriou, D., 2018. A systematic review of augmented reality content-related techniques for knowledge transfer in maintenance applications. Comput. Ind. 103, 47–71.

Deore, H., Agrawal, A., Jaglan, V., Nagpal, P., Sharma, M.M., 2020. A new approach for navigation and traffic signs indication using map integrated augmented reality for self-driving cars. Scalable Comput.: Pract. Exp. 21 (3), 441–450.

Dirin, A., Alamäki, A., Suomala, J., 2019. Gender differences in perceptions of conventional video, virtual reality and augmented reality. Int. Assoc. Online Eng. 13 (6), 93–1 03.

Eckert, M., Blex, M., Friedrich, C.M. , 2018, January. Object detection featuring 3D audio localization for Microsoft HoloLens. In: Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies. Vol. 5, pp. 555–561.

Farasin, A., Peciarolo, F., Grangetto, M., Gianaria, E., Garza, P., 2020. Real-time object detection and tracking in mixed reality using microsoft hololens. Proceedings of the15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2020, vol. 4. SciTePress, pp. 165–172.

Fischler, M.A., Elschlager, R.A., 1973. The representation and matching of pictorial structures. IEEE Trans. Comput. 100 (1), 67–92.

Fuchs, K., Grundmann, T., Fleisch, E. , 2019, October. Towards identification of packaged products via computer vision: Convolutional neural networks for object detection and image classification in retail environments. In: Proceedings of the 9th International Conference on the Internet of Things. pp. 1–8.

Fuchs, K., Haldimann, M., Grundmann, T., Fleisch, E., 2020. Supporting food choices in the internet of people: automatic detection of diet-related activities and display of real-time interventions via mixed reality headsets. Future Gener. Comput. Syst. 113, 343–362.

Geng, J., Song, X., Pan, Y., Tang, J., Liu, Y., Zhao, D., Ma, Y., 2020. A systematic design method of adaptive augmented reality work instruction for complex industrial operations. Comput. Ind. 119, 103229.

Girshick, R. , 2015. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1440–1448.

Golnari, A., Khosravi, H., Sanei, S. , 2020, February. DeepFaceAR: deep face recognition and displaying personal information via augmented reality. In: Proceedings of the 2020 International Conference on Machine Vision and Image Processing (MVIP). pp. 1–7. IEEE.

Han, L., Zheng, T., Zhu, Y., Xu, L., Fang, L., 2020. Live semantic 3D perception for immersive augmented reality. IEEE Trans. Vis. Comput. Graph. 26 (5), 2012–2022.

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.

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.

Hidalgo, M., Harris, S., Boland, W., Halfman, T., Johnston, J., Hillyer, T., Elliott, L., 2021. Training capabilities assessment in support of enhanced military training: comparing head-mounted displays. Proceedings of the International Conference on Applied Human Factors and Ergonomics. Springer, Cham, pp. 11–18.

Hu, M., Weng, D., Chen, F., Wang, Y. , 2020, October. Object detecting augmented reality system. In: Proceedings of the 2020 IEEE 20th International Conference on Communication Technology (ICCT). pp. 1432–1438. IEEE.

Huang, S., Han, T., Xie, J. , 2019, December. A smart-decision system for realtime mobile AR applications. In: Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM). pp. 1–6. IEEE.

Huynh, B., Orlosky, J., Höllerer, T. , 2019, March. In-situ labeling for augmented reality language learning. In: Proceedings of the 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). pp. 1606–1611. IEEE.

Jain, D.K., 2019. An evaluation of deep learning based object detection strategies for threat object detection in baggage security imagery. Pattern Recognit. Lett. 120, 112–119.

Karambakhsh, A., Kamel, A., Sheng, B., Li, P., Yang, P., Feng, D.D., 2019. Deep gesture interaction for augmented anatomy learning. Int. J. Inf. Manag. 45, 328–336.

Kästner, L., Frasineanu, V.C., Lambrecht, J., 2020. May. A 3D-deep-learning-based augmented reality calibration method for robotic environments using depth sensor data. Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 1135–1141.

Kästner, L., Eversberg, L., Mursa, M., Lambrecht, J. , 2021, January. Integrative object and pose to task detection for an augmented-reality-based human assistance system using neural networks. In: Proceedings of the 2020 IEEE Eighth International Conference on Communications and Electronics (ICCE). pp. 332–337. IEEE.

Katiyar, A., Kalra, K., Garg, C., 2015. Marker based augmented reality. Adv. Comput. Sci. Inf. Technol. (ACSIT) 2 (5), 441–445.

Khan, N., Saleem, M.R., Lee, D., Park, M.W., Park, C., 2021. Utilizing safety rule correlation for mobile scaffolds monitoring leveraging deep convolution neural networks. Comput. Ind. 129, 103448.

Kim, Y.H., Lee, K.H., 2019. Pose initialization method of mixed reality system for inspection using convolutional neural network. J. Adv. Mech. Des. Syst., Manuf. 13 (5) JAMDSM0093-JAMDSM0093.

Konstantinidis, F.K., Kansizoglou, I., Santavas, N., Mouroutsos, S.G., Gasteratos, A., 2020. MARMA: a mobile augmented reality maintenance assistant for fast-track repair procedures in the context of industry 4.0. Machines 8 (4), 88.

Lai, Z.H., Tao, W., Leu, M.C., Yin, Z., 2020. Smart augmented reality instructional system for mechanical assembly towards worker-centered intelligent manufacturing. J. Manuf. Syst. 55, 69–81.

Lang, Y., Liang, W., Yu, L.F. , 2019, March. Virtual agent positioning driven by scene semantics in mixed reality. In: Proceedings of the 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). pp. 767–775. IEEE.

Le, H., Nguyen, M., Yan, W.Q. , 2020, November. Machine learning with synthetic data – a new way to learn and classify the pictorial augmented reality markers in real-time. In: Proceedings of the 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ). pp. 1–6. IEEE.

Li, C., Sun, X., Li, Y., 2019. Information hiding based on augmented reality. Math. Biosci. Eng. 16 (5), 4777–4787.

Li, X., Tian, Y., Zhang, F., Quan, S., Xu, Y. , 2020, November. Object detection in the context of mobile augmented reality. In: Proceedings of the 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp. 156163. IEEE.

Li, Y., Wang, H., Dang, L.M., Nguyen, T.N., Han, D., Lee, A., Moon, H., 2020. A deep learning-based hybrid framework for object detection and recognition in autonomous driving. IEEE Access 8, 194228–194239.

Lin, C.H., Chung, Y., Chou, B.Y., Chen, H.Y., Tsai, C.Y. , 2018, April. A novel campus navigation APP with augmented reality and deep learning. In: Proceedings of the 2018 IEEE International Conference on Applied System Invention (ICASI). pp. 1075–1077. IEEE.

Lin, H.C., Wu, Y.H., 2017. Augmented reality using holographic display. Opt. Data Process. Storage 3 (1), 101–106.

Liu, L., Li, H., Gruteser, M. , 2019, August. Edge assisted real-time object detection for mobile augmented reality. In: Proceedings of the The 25th Annual International Conference on Mobile Computing and Networking. pp. 1–16.

Liu, Q., Han, T. , 2018, September. Dare: Dynamic adaptive mobile augmented reality with edge computing. In: Proceedings of the 2018 IEEE 26th International Conference on Network Protocols (ICNP). pp. 1–11. IEEE.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. Ssd: Single shot multibox detector. Proceedings of the European Conference on Computer Vision. Springer, Cham, pp. 21–37.

Livingston, M.A., Rosenblum, L.J., Brown, D.G., Schmidt, G.S., Julier, S.J., Baillot, Y., Maassel, P., 2011. Military applications of augmented reality. Handb. Augment. Real. 671–706.

Llasag, R., Marcillo, D., Grilo, C., Silva, C. , 2019, June. Human detection for search and rescue applications with uavs and mixed reality interfaces. In: Proceedings of the 2019 14th Iberian Conference on Information Systems and Technologies (CISTI). pp. 1–6. IEEE.

Lomaliza, J.P., Park, H., 2020. Initial pose estimation of 3D object with severe occlusion using deep learning. Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems. Springer, Cham, pp. 325–336.

Mahurkar, S. , 2018, November. Integrating YOLO object detection with augmented reality for iOS Apps. In: Proceedings of the 2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). pp. 585–589. IEEE.

McKelvey, C., Dreyer, R., Zhu, D., Wang, W., Quarles, J. , 2019, October. Energy-oriented designs of an augmented-reality application on a VUZIX blade smart glass. In: Proceedings of the 2019 Tenth International Green and Sustainable Computing Conference (IGSC). pp. 1–8. IEEE.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., 2010. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int. J. Surg. 8 (5), 336–341.

Nilwong, S., Capi, G. , 2020, June. Outdoor robot navigation system using game-based DQN and augmented reality. In: Proceedings of the 2020 17th International Conference on Ubiquitous Robots (UR). pp. 74–80. IEEE.

Pai, N.S., Huang, J.B., Wu, J.X., Chen, P.Y., Zhou, Y.H., 2020. Forward collision warning and lanemark recognition systems based on deep learning. Sens. Mater. 32 (6), 1981–1995.

Park, K.B., Choi, S.H., Kim, M., Lee, J.Y., 2020a. Deep learning-based mobile augmented reality for task assistance using 3D spatial mapping and snapshot-based RGB-D data. Comput. Ind. Eng. 146, 106585.

Park, K.B., Kim, M., Choi, S.H., Lee, J.Y., 2020b. Deep learning-based smart task assistance in wearable augmented reality. Robot. Comput. -Integr. Manuf. 63, 101887.

Park, K.B., Choi, S.H., Lee, J.Y., Ghasemi, Y., Mohammed, M., Jeong, H., 2021. Hands-free human–robot interaction using multimodal gestures and deep learning in wearable mixed reality. IEEE Access 9, 5544855464.

Park, Y.J., Ro, H., Lee, N.K., Han, T.D., 2019. Deep-care: projection-based home care augmented reality system with deep learning for elderly. Appl. Sci. 9 (18), 3897.

Plecher, D.A., Eichhorn, C., Seyam, K.M., Klinker, G. , 2020, November. A Rsinoëlearning egyptian hieroglyphs with augmented reality and machine learning. In: Proceedings of the 2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). pp. 326–332. IEEE.

Połap, D., Kęsik, K., Książek, K., Woźniak, M., 2017. Obstacle detection as a safety alert in augmented reality models by the use of deep learning techniques. Sensors 17 (12), 2803.

Putze, F., Küster, D., Urban, T., Zastrow, A., Kampen, M. , 2020, October. Attention sensing through multimodal user modeling in an augmented reality guessing game. In: Proceedings of the 2020 International Conference on Multimodal Interaction. pp. 33–40.

Ramakrishna, P., Hassan, E., Hebbalaguppe, R., Sharma, M., Gupta, G., Vig, L., Shroff, G. , 2016, September. An ar inspection framework: Feasibility study with multiple ar devices. In: Proceedings of the 2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct). pp. 221–226. IEEE.

Ran, X., Chen, H., Liu, Z., Chen, J. , 2017, August. Delivering deep learning to mobile devices via offloading. In: Proceedings of the Workshop on Virtual Reality and Augmented Reality Network. pp. 42–47.

Rao, J., Qiao, Y., Ren, F., Wang, J., Du, Q., 2017. A mobile outdoor augmented reality method combining deep learning object detection and spatial relationships for geovisualization. Sensors 17 (9), 1951.

Rathnayake, D., de Silva, A., Puwakdandawa, D., Meegahapola, L., Misra, A., Perera, I. , 2020, December. Jointly optimizing sensing pipelines for multimodal mixed reality interaction. In: Proceedings of the 2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS). pp. 309–317. IEEE.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A. , 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28, 91–99.

Rodrigues, J.M., Veiga, R.J., Bajireanu, R., Lam, R., Pereira, J.A., Sardo, J.D., Bica, P., 2018. Mobile augmented reality framework-MIRAR. Proceedings of the International Conference on Universal Access in Human Computer Interaction. Springer, Cham, pp. 102–121.

Rosenfeld, A., Pfaltz, J.L., 1966. Sequential operations in digital picture processing. J. ACM 13 (4), 471–494.

Sharma, V., Mir, R.N., 2020. A comprehensive and systematic look up into deep learning based object detection techniques: a review. Comput. Sci. Rev. 38, 100301.

Singh, A., Ghasemi, Y., Jeong, H., Kim, M., Johnson, A., 2021. A comparative evaluation of the wearable augmented reality-based data presentation interface and traditional methods for data entry tasks. Int. J. Ind. Ergon. 86, 103190.

Subakti, H., Jiang, J.R. , 2018, July. Indoor augmented reality using deep learning for industry 4.0 smart factories. In: Proceedings of the 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC). Vol. 2, pp. 63–68. IEEE.

Sun, Y., Kantareddy, S.N. R., Siegel, J., Armengol-Urpi, A., Wu, X., Wang, H., Sarma, S., 2019. Towards industrial IOT-AR systems using deep learning-based object pose estimation. Proceedings of the 2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC). IEEE, pp. 1–8.

Sutanto, R.E., Pribadi, L., Lee, S., 2017. 3D integral imaging based augmented reality with deep learning implemented by faster R-CNN. Proceedings of the International Conference on Mobile and Wireless Technology. Springer, Singapore.

Syed, A.T., Merugu, S., Kumar, V., 2020. Augmented reality on sudoku puzzle using computer vision and deep learning. Advances in Cybernetics, Cognition, and Machine Learning for Communication Technologies. Springer, Singapore, pp. 567–578.

Tao, W., Lai, Z.H., Leu, M.C., Yin, Z., Qin, R., 2019. A self-aware and active-guiding training & assistant system for worker-centered intelligent manufacturing. Manuf. Lett. 21, 45–49.

Tobías, L., Ducournau, A., Rousseau, F., Mercier, G., Fablet, R. , 2016, December. Convolutional neural networks for object recognition on mobile devices: a case study. In: Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR). pp. 3530–3535. IEEE.

Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W., 2013. Selective search for object recognition. Int. J. Comput. Vis. 104 (2), 154–171.

Waithe, D., Brown, J.M., Reglinski, K., Diez-Sevilla, I., Roberts, D., Eggeling, C., 2020. Object detection networks and augmented reality for cellular detection in fluorescence microscopy. J. Cell Biol. 219 (10), e201903166.

Wang, J., Xiao, R., Jia, L., Wang, X., 2019. Mixed reality medical first aid training system based on body identification. Proceedings of the International Conference on Image and Graphics. Springer, Cham, pp. 395–406.

Wang, R., Lu, H., Xiao, J., Li, Y., Qiu, Q. , 2018, August. The design of an augmented reality system for urban search and rescue. In: Proceedings of the 2018 IEEE International Conference on Intelligence and Safety for Robotics (ISR). pp. 267–272. IEEE.

Wang, S., Guo, R., Wang, H., Ma, Y., Zong, Z. , 2018, August. Manufacture assembly fault detection method based on deep learning and mixed reality. In: Proceedings of the 2018 IEEE International Conference on Information and Automation (ICIA). pp. 808–813. IEEE.

Zamora-Hernández, M.A., Castro-Vargas, J.A., Azorin-Lopez, J., Garcia-Rodriguez, J., 2021. Deep learningbased visual control assistant for assembly in industry 4.0. Comput. Ind. 131, 103485.

Zheng, L., Liu, X., An, Z., Li, S., Zhang, R., 2020. A smart assistance system for cable assembly by combining wearable augmented reality with portable visual inspection. Virtual Real. Intell. Hardw. 2 (1), 12–27.

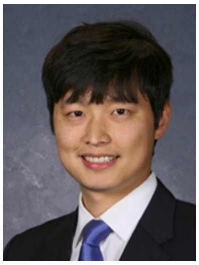
Zhou, P., Braud, T., Zavodovski, A., Liu, Z., Chen, X., Hui, P., Kangasharju, J., 2020. Edge-facilitated augmented vision in vehicle-to-everything networks. IEEE Trans. Veh. Technol. 69 (10), 1218712201.

Zhou, Z., Li, L., Fürsterling, A., Durocher, H.J., Mouridsen, J., Zhang, X., 2022. Learning-based object detection and localization for a mobile robot manipulator in SME production. Robot. Comput. -Integr. Manuf. 73, 102229.

Židek, K., Hosovsky, A., Piteľ, J., Bednár, S., 2019a. Recognition of assembly parts by convolutional neural networks. Advances in Manufacturing Engineering and Materials. Springer, Cham, pp. 281–289.

Židek, K., Lazorík, P., Piteľ, J., Hošovský, A., 2019b. An automated training of deep learning networks by 3D virtual models for object recognition. Symmetry 11 (4), 496.

Yalda Ghasemi received her B.S. degree in industrial engineering from the Shomal University, Iran. She is currently a Ph.D. candidate in industrial engineering and operations research at the University of Illinois at Chicago. Her current research interests include human-computer interaction and human factors in extended reality applications.



Heejin Jeong received his B.S. degree in industrial engineering from the Pohang University of Science and Technology, South Korea, in 2010, and the M.S.E. and Ph.D. degrees in industrial and operations engineering from the University of Michigan, Ann Arbor, in 2018. He is currently an Assistant Professor in the Department of Mechanical and Industrial Engineering, University of Illinois at Chicago. His current research interests include human factors engineering, cognitive ergonomics, and human performance modeling.

Sung Ho Choi received his BS and MS degrees in Industrial Engineering from Chonnam National University, South Korea. He is currently a Ph.D. candidate at Chonnam National University. His current research interests include AR-based remote collaboration and human-robot collaboration

Kyeong-Beom Park received his BS and MS degrees in Industrial Engineering from Chonnam National University, South Korea. He is currently a Ph.D. candidate at Chonnam National University. His current research interests include AR/MR and deep learning-based applications.

Jae Yeol Lee is a professor in the Department of Industrial Engineering, Chonnam National University, South Korea. He received his BS, MS and Ph.D. degrees in Industrial Engineering from Pohang University of Science and Technology (POSTECH), South Korea, in 1992, 1994, and 1998, respectively. From 1998–2003, he worked as a senior researcher at Electronics and Telecommunications Research Institute (ETRI), Korea. Since 2003, he has been a faculty member of Chonnam National University. His current research interests include AR/MR, deep learning, human-robot collaboration, and UX.