IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 January 2024, accepted 22 February 2024, date of publication 27 February 2024, date of current version 15 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3371022

RESEARCH ARTICLE

People Identification in Private Car Using 3D LiDAR With Generative Image Inpainting and YOLOv5

WEIRONG SHAO[1], (Graduate Student Member, IEEE),
MONDHER BOUAZIZI[2], (Member, IEEE),
XIANG MENG[1], (Graduate Student Member, IEEE),
AND TOMOAKI OHTSUKI[2], (Senior Member, IEEE)
[1]Graduate School of Science and Technology, Keio University, Yokohama 223-8522, Japan
[2]Faculty of Science and Technology, Keio University, Yokohama 223-8522, Japan

Corresponding author: Tomoaki Ohtsuki (ohtsuki@ics.keio.ac.jp)

This work was supported by JST ASPIRE, Japan, under Grant JPMJAP2326.

ABSTRACT People Identification is a critical aspect in developing modern vehicles, aimed at enhancing safety and comfort levels. Most traditional methods of people identification in vehicles use RGB images or videos. In this study, we introduce a novel methodology for identifying individuals in private car scenarios, utilizing 3D Light Detection and Ranging (LiDAR) technology, generative image inpainting based on Contextual Attention, and the YOLOv5 model. Initially, we gather data utilizing a 3D-LiDAR instrument and subsequently convert the acquired depth data into depth images. Following this, the depth images are annotated manually to indicate the positions and identifiers of various individuals occupying distinct seats. This annotated data serves as the training material for the YOLOv5 model, facilitating the recognition and categorization of subjects. However, given that individuals seated in the back often have parts of their bodies occluded by the front seats and the passengers in them, we employ generative image inpainting techniques to reveal the occluded portions. This step significantly enhances the precision in detecting and identifying individuals situated in the back seats. We implemented our strategy on a restricted group of four participants, conducting training and testing phases within identical environments. Prior to the inpainting process, the classification's F1 score stood at 66.5%. After inpainting, we observed a notable surge in the F1 score for the rear-seat passengers increased by 17.1%.

INDEX TERMS YOLOv5, GAN, people identification, 3D LiDAR, deep learning.

I. INTRODUCTION

In the swiftly evolving landscape of modern society, automobiles have become one of the most important tools in every family. It greatly improves the convenience of every family's life. In the meantime, people identification technologies have developed rapidly and are widely implemented in vehicles. With the use of this technology, human behavior while driving and human identity can be more accurately recognized, improving overall driving comfort and safety. Human identity can be observed, learned from, and predicted using in-vehicle person recognition technology. It has the ability to assess the driver's degree of focus, spot indicators of drowsiness, warn the driver or even take over driving to prevent accidents [1]. Additionally, it can recognize the driver and adjust the vehicle's settings in accordance with their preferences, which adds another level of comfort and customization [2].

Current methods for identifying individuals predominantly utilize RGB images or video footage [3]. Such data can be compromised by environmental conditions, such as varying lighting or unfavorable weather, leading to potential

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

inaccuracies in identification. Beyond these challenges, there are also significant privacy concerns associated with the use of RGB images and videos, as they can capture detailed personal information [4]. Given these limitations, it's crucial to explore alternative approaches for people recognition that are both resilient to environmental variables and more protective of individual privacy. Such innovations could pave the way for safer and more reliable identification systems in various applications.

To address the inherent issues in contemporary people identification practices, 3D Light Detection and Ranging (LiDAR) technology is used in our experiment. Our experiment is improving the limitations of traditional RGB images by utilizing infrared laser beams. These beams meticulously map out the surrounding environment, generating a detailed 3D spatial representation. Through this method, 3D LiDAR ensures the acquisition of precise distance and angle data, markedly more accurate than what RGB images can offer.

3D LiDAR is distinctly advantageous in safeguarding privacy [4]. Whereas RGB cameras capture detailed images that may inadvertently reveal personal information, 3D LiDAR gracefully sidesteps this problem by recording only the general outlines of objects and distance information within its view. This approach ensures that personal details remain confidential, striking a balance between identification needs and privacy preservation.

Another feature of 3D LiDAR technology is its great performance under various lighting conditions, including complete darkness. This characteristic is particularly invaluable for applications such as people identification within private cars, where the interior lighting is often minimal or non-existent, especially during the night. In such scenarios, 3D LiDAR not only maintains its effectiveness but proves to be a superior alternative, capable of functioning optimally in low-light conditions and protecting passenger privacy simultaneously. Therefore, employing 3D LiDAR devices promises a robust solution for people identification across diverse environments, addressing and resolving the challenges posed by traditional identification technologies.

Despite its strengths, employing 3D LiDAR for people identification within private car settings is not without drawbacks. A primary challenge arises when attempting to detect passengers seated in the back row. More often than not, the data representing these passengers is occluded by the front seats and their occupants. This occlusion leads to suboptimal detection precision and hinders the overall accuracy of classifying individuals.

To solve this challenge, our experiment involves utilizing a generative image inpainting. Generative image inpainting is designed to restore and repair sections of an image that might be missing or compromised [5]. This technique proves useful as it aids in reconstructing the occluded portions of the bodies of rear passengers. By doing so, it significantly boosts the precision in detection, ensuring a more accurate identification process. Through the integration of this method, we envision a holistic system that capitalizes on the strengths of 3D LiDAR while simultaneously mitigating its limitations in private car contexts.

To summarize, our main contributions are as follows:

1) Transfer depth data captured by 3D LiDAR device into depth images.
2) Reconstruct the people's occluded portions from the back seat.
3) Validate the potential of the YOLOv5 model applied to people identification with 3D-LiDAR images and compare the result before and after reconstruction.

Our early experimental results have been submitted as a conference manuscript for publication in [6]. The current manuscript explains in more details our proposed method, elaborates on the tuning of the parameters of our models and experimental settings, and gives more robust basis for our choices of models, parameters and scenarios. Nonetheless, more detailed results are given with a more thorough analysis of the results, the merits and limitations of our proposed method, and the potential future challenges that need to be addressed.

In our experiment, we utilize Dynamic Non-linear Mapping (DNLM) [7] to transform depth information into depth images. This transformation provides the foundation for more sophisticated visual analyses.

A significant challenge we encountered pertained to the accurate identification of people seated in the back seats of private cars. Due to the inherent limitations of traditional methods and the occasional obstructions in the visual field, identifying these individuals with clarity may be difficult. Addressing this problem, we devised a method to reconstruct the occluded part of the body, significantly enhancing the resolution and clarity of these visual representations.

The next part of our research pivoted around the evaluation of the YOLOv5 model [8], which is the model utilized in the field of object detection. Known for its robust capabilities, we sought to understand its performance specifically in the realm of 3D-LiDAR-based people identification. Our experiments confirmed its exemplary potential in this domain. Moreover, we contrast the identification results before and after the image reconstruction to evaluate the role of generative image inpainting in people identification. These findings underscored the profound impact of our reconstruction technique when combined with YOLOv5.

In conclusion, the novelty of our work is that it combines 3D LiDAR technology with generative image inpainting algorithms to drastically improve recognition accuracy in these situations, which solves the drawback of RGB images and videos that cannot be clear when the environment is dark. 3D LiDAR provides precise depth information and spatial awareness to effectively identify and locate people in vehicles, making accurate judgments even when part of the line of sight is obscured. Additionally, we employ DNLM technology to enhance the clarity of depth images, particularly in confined spaces such as the interior of private cars.

This approach ensures that even in these narrow scenarios, the depth images are distinctly rendered, facilitating more accurate analysis and interpretation. Notably, there is no existing state-of-the-art research that has utilized 3D LiDAR for individual identification in such contexts.

II. RELATED WORK AND MOTIVATIONS

A. MOTIVATIONS AND CHALLENGES

3D LiDAR (Light Detection and Ranging) has emerged as a promising technology for human identification due to its ability to capture high-resolution spatial information. One of the primary motivations behind its adoption is its capability to produce detailed 3D point clouds [9] and depth information [10], which can reveal nuances in human shapes and gaits. Moreover, unlike traditional camera systems, LiDAR operates independently of lighting conditions, making it an ideal choice for environments with variable or low lighting. This feature not only enhances its versatility but also respects personal privacy, as LiDAR does not capture facial features or other sensitive personal details in the same way cameras do [11]. The integration of 3D LiDAR with other sensor systems, such as RGB cameras [12], further augments its accuracy, creating a robust multi-model identification system.

However, the use of 3D LiDAR in human identification is not without challenges. One of the primary hurdles is the current lack of established identification methods tailored to LiDAR data, given that traditional biometrics like facial recognition are more mature [13]. Human figures, being inherently variable in shape, size, and posture, introduce additional challenges in ensuring consistent identification. The data collected can also be affected by noise and environmental artifacts, especially in indoor settings, where people may be partially occluded, complicating the identification process, and making feature extraction a complex task.

B. RELATED WORK

3D LiDAR is a form of light detection and ranging that produces three-dimensional (3D) maps by measuring the time it takes for laser beams to reflect back after being emitted [14]. The use of 3D LiDAR has been widespread in several fields. One of the most basic applications of 3D LiDAR for humans is to detect and track individuals in various environments [15], [16], [17]. Yan et al. [18], [19] presented that 3D LiDAR can create a high-resolution point cloud, which can be used to discern the general shape and size of a human. In scenarios such as public events or busy transportation hubs, LiDAR can provide insights into crowd dynamics, enabling efficient crowd management and safety precautions [20]. Lin et al. utilized 3D LiDAR in the field of place recognition [21]. It is conducive to application in the field of robotics.

Gait, or the way an individual walks, is unique and can serve as a biometric identifier. With LiDAR's detailed spatial data, researchers have been able to study the subtle nuances in gait patterns [22], [23]. Such patterns can be used for identifying individuals or detecting abnormalities in walking patterns, potentially useful for medical applications. In the context of fall detection, especially relevant in healthcare and elderly care, LiDAR can be employed to detect if a person has fallen, enabling timely medical intervention [24], [25].

The fusion of machine learning, particularly deep learning, with 3D LiDAR data has expanded the potential for people identification. Techniques such as Convolutional Neural Networks (CNNs) process 3D point cloud data [26], [27], enabling intricate feature extraction and recognition. Specifically, architectures like PointNet have been tailored to directly handle point cloud data, enhancing the accuracy and reliability of identification systems [28]. Zhou and Tuzel [29] presented an architecture called VoxelNet for point clouds. VoxelNet extracts features from sparse points on the 3D voxel grid and obtains outstanding performance on the KITTI benchmark dataset.

Traditional object detection is a two-stage algorithm. Pioneering the realm of two-stage detectors, Girshick et al. [30] introduced the Region-based Convolutional Neural Network (R-CNN). The R-CNN first employed selective search to generate around 2000 region proposals, which were then classified using CNNs. Although achieving state-of-the-art performance, R-CNN is computationally expensive and unsuitable for real-time applications. Addressing the computational inefficiencies of its predecessor, Girshick proposed Fast R-CNN [31]. Unlike R-CNN, which applied the CNN on each proposed region separately, Fast R-CNN processed the entire image with a CNN first and then extracted features for each region proposal, significantly boosting its speed. Extending the capabilities of Faster R-CNN, He et al. introduced Mask R-CNN [32], which added a parallel branch for predicting segmentation masks. This enhancement made it adept at instance segmentation tasks alongside object detection.

YOLO [33] is a real-time object detection system based on deep learning. YOLO employs a one-shot method to object detection, which significantly enhances the detection speed and lets it perform well in real-time applications, in contrast to other object detection algorithms like Fast R-CNN [31], [34]. The convolutional neural network (CNN) used by YOLO receives the complete image and outputs object class and location data in a single forward propagation. It has been used in the realm of 3D LiDAR technology. YOLO is used in object classification and detection combined with LiDAR and camera fusion [35], [36], [37]. Besides that, Simon et al. [38], [39], [40] proposed a contribution to the field of 3D object detection with Complex-YOLO. This neural network architecture is meticulously crafted for real-time 3D object detection tasks. Unlike traditional models which might rely on a combination of data types, Complex-YOLO stands out as it operates directly on point cloud data. Furthermore, its adaptability extends to semantic point cloud data, both of which are generated by 3D LiDAR systems [41]. These synergies between deep learning and 3D LiDAR have

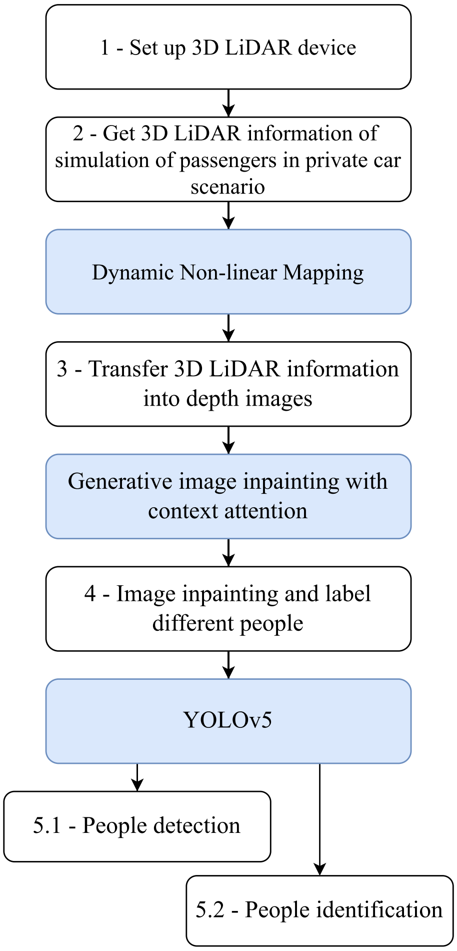

1 - Set up 3D LiDAR device
2 - Get 3D LiDAR information of simulation of passengers in private car scenario
Dynamic Non-linear Mapping
3 - Transfer 3D LiDAR information into depth images
Generative image inpainting with context attention
4 - Image inpainting and label different people
YOLOv5
5.1 - People detection
5.2 - People identification

FIGURE 1. A flowchart describing the steps of our proposed framework.

paved the way for more sophisticated and accurate people identification solutions. The combination of YOLO and 3D LiDAR has gained widespread recognition in the autonomous fields [42]. This integration has been pivotal in advancing autonomous object detection, as exemplified by the work of Tian and Guo [43]propose a novel approach combining YOLO with 3D LiDAR. Further research into this domain, Wu et al. [44] have centered their research on the on-road detection of objects using both camera and 3D LiDAR, offering insights into regional aspects of object detection in autonomous systems.

III. OVERALL SYSTEM DESCRIPTION AND EQUIPMENT

A. OVERALL SYSTEM DESCRIPTION

The flow of our experiment is given in Figure 1. Initially, depth data is captured by a 3D LiDAR device. This data is then transformed into depth images via Dynamic Non-linear Mapping (DNLM) [7]. To enhance the image quality, any missing parts are reconstructed using generative image inpainting with contextual attention [45], [46].

Following this, individuals in the images are annotated. The annotated images serve as training data for the YOLOv5 model, optimizing it for people detection from LiDAR images. The trained model is then proficiently employed for the detection and identification of people in private car scenarios.

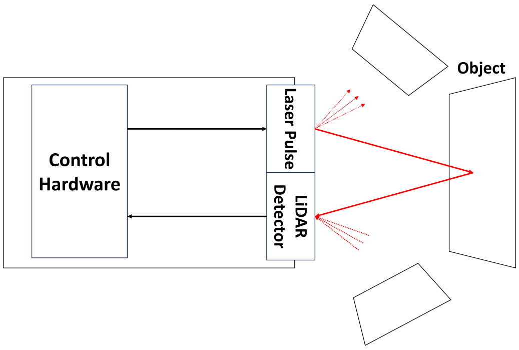


FIGURE 2. LiDAR system.

B. EQUIPMENT

LiDAR (Light Detection and Ranging) is a remote sensing technology that uses infrared laser (IR) beams to measure the distance between an object and the sensor. Traditional RGB cameras require visible light to record images, in contrast to 3D LiDAR, which can function well even in complete darkness. 3D LiDAR specifically refers to devices capable of capturing spatial information in three dimensions. The working principle is as follows: Initially, the system emits a brief, high-precision laser pulse. When this pulse encounters an object's surface, it gets reflected and is captured by the LiDAR detector. By calculating the time difference between the emission and reception of the laser pulse, and considering the speed of light, the system can accurately measure the distance to the object. The distance is calculated according to Eq. (1):

$$R = C \cdot \frac{T}{2}, \tag{1}$$

where R is the measured distance, C is the speed of light, and T is the time difference between the laser signal emission and reception.

To acquire three-dimensional data, the LiDAR scans in multiple directions and angles, typically achieved through mechanical rotation, galvanometer scanning, or optical phased arrays, the system is shown in Figure 2. Ultimately, this data is transformed into a three-dimensional point cloud model, where each point possesses its x, y, and z coordinates in space, which can be seen in Figure 3, and the coordinates are obtained from the Eqs. (2) to (4),

$$x = S \cdot \sin\phi \cdot \cos\theta, \tag{2}$$

$$y = S \cdot \sin\phi \cdot \sin\theta, \tag{3}$$

$$z = S \cdot \cos\phi, \tag{4}$$

where S is the distance measured by LiDAR, ϕ is the vertical scanning angle of the laser pulse, and θ is the horizontal scanning angle.

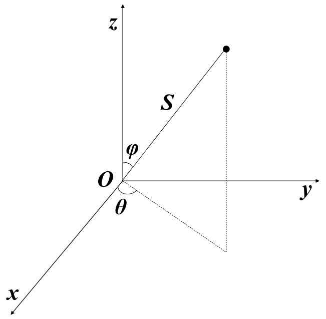


FIGURE 3. LiDAR three-dimensional coordinate.

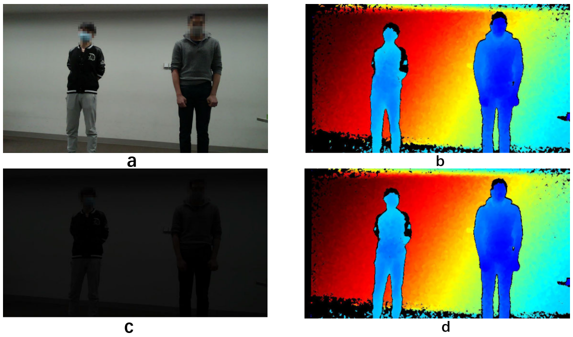


FIGURE 4. 3D LiDAR generated depth images in different illuminations.

By rapidly repeating this process, the 3D information of the environment being surveyed may be generated using 3D LiDAR. By collecting both the distance and the angle of each item in reference to the sensor, this representation gives a highly detailed picture of the environment. When setting up the 3D LiDAR equipment, we verified the imaging status of depth images under various lighting conditions. Figure 4(a) presents an RGB image of two experimental participants in a bright environment, while Figure 4(b) displays the corresponding depth image. Figure 4(c) is an RGB image of two participants in a relatively dark environment, and Figure 4(d) is its corresponding depth image. It can be seen that there is no significant difference between the generated depth images under different lighting conditions. It proves that 3D LiDAR can perform well in different lighting conditions.

DNLM is the technique that transfers 3D LiDAR information into 3D LiDAR images, which are also called depth images. It helps to accurately capture and interpret complex nonlinear relationships in raw 3D data, ensuring the authenticity and detail of 3D images.

In our experiment, we use a 3D+RGB IP67 Kit (Helios2 & Triton 3.2MP Kit), the specifications of this device are given in TABLE 1. We set up experiments using the LUCID Arena SDK, provided by LUCID Vision Labs, to collect 3D LiDAR information. The details of LUCID Arena SDK are provided in TABLE 2. We use OpenCV to perform image processing tasks such as generating and saving depth images; Pytorch libraries to invoke, fine-tune, and infer the necessary deep-learning models for people identification and image inpainting. Data are collected as separate scenarios which simulate private cars. The device is placed in front of four passengers. The distance between the device and the front passengers is less than 1.0 m and the rear passengers are about 1.5 m which is catering to the scene in a real private car.

TABLE 1. The specifications of the 3D LiDAR used in the experiments.

Parameter	Value
Number of pixels	0.3 MP
Resolution	640 × 480 px
Frame rate	30 fps
Angle of view	69° × 51°
VCSEL wavelength	850 nm, Indoors
Number of exposure modes	62.5/250/1000 μs

TABLE 2. LUCID Arena SDK for windows.

Feature	Description
Data Interface	Gigabit Ethernet (1000 Mbit/s)
Operating System	Windows 7 / 10 (32-bit and 64-bit)
Compiler	Visual Studio 2015 project files included.
Programming Languages	C++, C, and C#

IV. DETAIL SYSTEM DESCRIPTION

A. DATA COLLECTION

In our experiment, we set three different types of scenarios:

1: In our first scenarios, we position four individuals, conveniently labeled as A, B, C, and D, in distinct seats facing a 3D LiDAR device. These subjects serve as the primary entities to validate the effectiveness of our proposed technique. To ensure comprehensive data capture, we rotate these subjects, making each one occupy different seats in successive intervals. In every such interval, approximately 500 frames of data are gathered as the subjects engage in a variety of activities like talking, playing mobile phones, and so on, culminating in a total of 2000 frames amassed specifically for this configuration.

2: Advancing to the next set of scenarios, we introduce an additional element of complexity by integrating an unknown individual into the mix. This unknown person sporadically takes the place of one of our original subjects (A, B, C, or D) from the first scenario. As with the previous setup, 2000 frames are collected for each iteration. Moreover, to maintain consistency, every unique seating permutation of the subjects results in 500 frames.

3: About the third set of scenarios, we aim to collect data that is used for image inpainting. Here, our primary subjects A, B, C, and D are directed to sequentially occupy two specific seats: the back left and the back right. This configuration is created to serve as a ground truth for our generative image inpainting tests.

After collecting multiple frames of 3D LiDAR data, the subsequent step involves transforming this 3D LiDAR

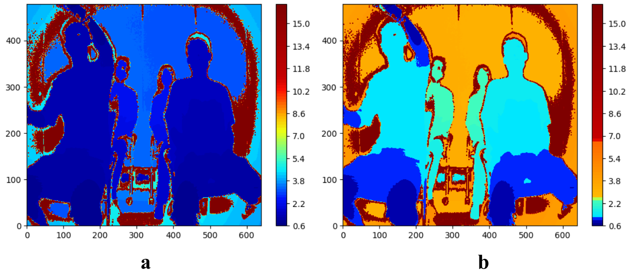


FIGURE 5. Depth images before and after using DNLM: Image a is before using DNLM, image b is after using DNLM.

information into depth images. A notable limitation of depth images is their representation of closely situated objects. If objects are in close proximity, their corresponding colors in the depth image tend to be similar. This similarity becomes particularly pronounced in private car settings where passengers are seated closely together. As a result, the depth-wise distinction between front and rear passengers frequently becomes inconspicuous, sometimes even blending indistinctly with the background.

To address this challenge, we've employed the advanced capabilities of DNLM to transfer 3D LiDAR data into depth images. DNLM, renowned for its dynamic adaptability, can be meticulously fine-tuned in real time. This versatility ensures that the depth images generated are tailored to fit a broad spectrum of scenarios, satisfying diverse requirements while making object differentiation significantly more effortless.

Illustratively, in Figure 5, we observe a color bar, referred to as 'bounds'. This bar is essentially a color map subdivided into several distinct sections. Each of these sections represents a specific color value, serving as a visual guide. Figure 5(**a**) seeks to establish a one-to-one mapping relationship between these bounds and chromatograms. In contrast, Figure 5(**b**) embarks on the mission of pinpointing the nonlinear center and finessing the bounds interval.

A keen observation of Figure 5 reveals that Figure 5(**a**) grapples with clarity issues, as it cannot distinguish between front passengers, rear passengers, and the background. This lack of clarity stems from the small distance separation between the front and rear seats. However, upon the application of DNLM to 3D LiDAR data, discerning between front and rear passengers becomes clear, cutting through the earlier ambiguity with precision. More examples of depth images generated by DNLM can be seen in Figure 6.

Such enhancement not only aids in immediate interpretation but also lays a robust foundation for more in-depth, subsequent research, paving the way for a better 3D LiDAR data analysis.

B. PEOPLE IDENTIFICATION MODEL

In recent times, advancements in deep learning methodologies coupled with the evolution of GPU hardware have greatly propelled the progress of computer vision technology. Using computer vision to minimize human labor has profound practical implications. Object detection is a fundamental aspect of digital image processing and computer vision. It also stands as the central component of intelligent surveillance systems across a wide range of applications. The YOLO object detection algorithm is the first single-stage object detection algorithm proposed by Jiang et al. [47]. It is an acronym that stands for You Only Look Once.

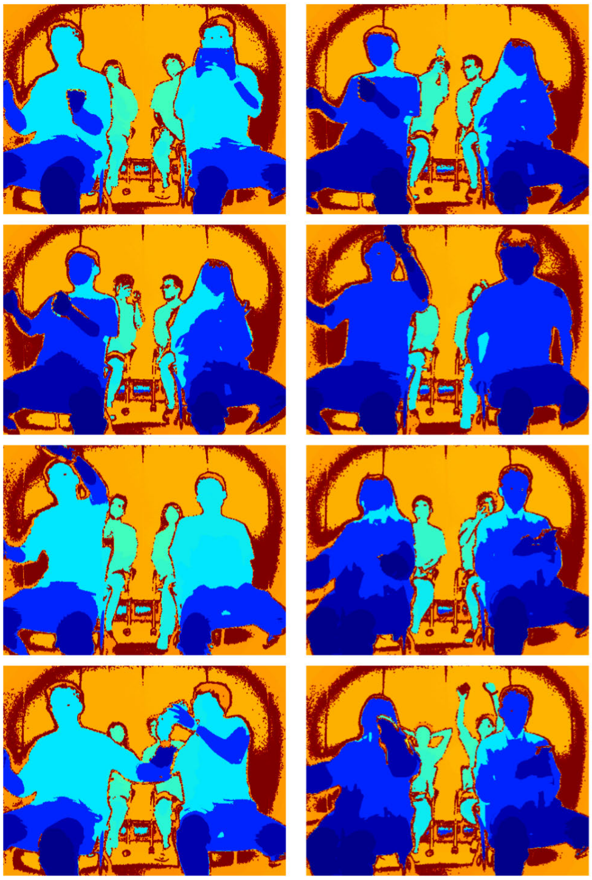


FIGURE 6. Examples of depth images.

The proposed method for people identification using 3D LiDAR relies on version 5 of YOLO (YOLOv5). It was introduced by Ultralytics in June 2020 [8]. Today, it stands as a prominent object detection algorithm. YOLOv5 is an innovative convolutional neural network (CNN) [48] adept at identifying objects in real-time with remarkable precision. Instead of multiple evaluations, this model scans the entire image using a single neural network pass. It then subdivides the image and forecasts bounding boxes and associated probabilities for each section. These bounding boxes get prioritized based on their predicted probabilities. The uniqueness of YOLO is in its method: it requires just a single look or one forward pass through the network to make its predictions. Post-prediction, detected objects are presented after a process called non-max suppression ensures each object is identified distinctly. YOLOv5 architecture is shown in Figure 7. From an architectural point of view, the YOLOv5 model is similar to YOLOv4 [49], and relies on three crucial sub-networks: the backbone, the neck, and the head.

- *Backbone:* The foundational element of any detection system, the backbone is responsible for feature extraction from the provided image. Within the framework, the utilization of CSPDarknet, which was proposed by Bochkovskiy et al. [49], emerges as a pivotal choice. Compared to Darknet53, which was prominently employed in the YOLOv3 model as indicated by Redmon and Farhadi [50], CSPDarknet presents a notable advancement. The core principle behind CSPDarknet is to split the feature map from the previous stage into two parts and then merge the partial features after a series of convolutions, which is the so-called Cross-Stage Hierarchical (CSH) feature [51]. This method not only ensures speed enhancements due to its streamlined architecture, but it also guarantees comparable, if not superior, detection accuracy. In essence, this optimized structure facilitates efficient and precise extraction of critical features from the input image.
- *Neck:* Acting as an intermediary between the backbone and the head, the neck has the primary function of managing and refining the features extracted by the backbone. Specifically, the Path Aggregation Network (PANet) [52] serves this purpose. It meticulously constructs feature pyramid networks (FPNs), which prove invaluable for the generalization across varying object scales. The fundamental advantage of PANet is its ability to guide information seamlessly from each extracted feature layer directly to the associated subnetwork. This ensures that no critical data is lost or diluted during the transition from the backbone to the subsequent stages.
- *Head:* The final part of the system, the head, undertakes the ultimate task of object detection. It produces anchor boxes corresponding to different feature maps, effectively serving as reference points for potential object locations. Moreover, the head outputs vectors that indicate class probabilities and the exact coordinates of detected bounding boxes. This operation aligns with the methodology employed in earlier YOLO iterations. In essence, the head transforms the abstract features procured and refined by the backbone and the neck into tangible and interpretable results.

According to the research of Ultralytics. YOLOv5 boasts significant improvements over most of the versions of YOLO in terms of size and speed while maintaining comparable accuracy. For example, YOLOv5 is approximately 88% smaller in size, being 27 MB compared to YOLOv4's 244 MB. Additionally, it operates at a speed that's around 180% faster, achieving 140 FPS compared to YOLOv4's 50 FPS. In terms of accuracy, both versions are closely matched on the same task.

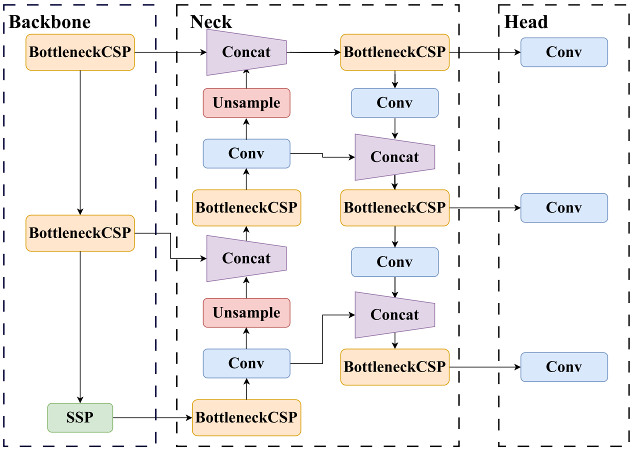
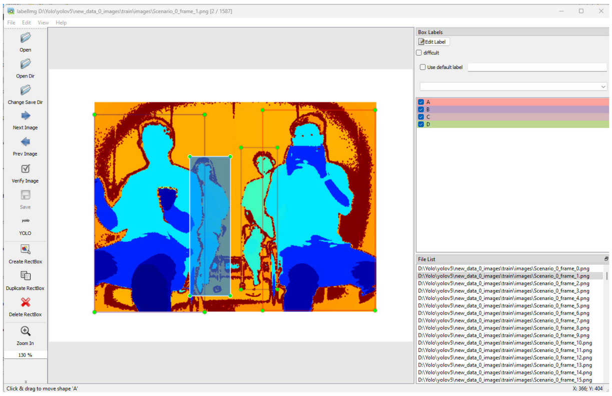


FIGURE 7. The architecture of YOLOv5 model. CSP stands for cross stage partial network. Conv stands for convolutional layer. SPP stands for spatial pyramid pooling. Concat stands for concatenate function.

TABLE 3. An example of YOLO format annotated data.

Subject ID	X center	Y center	Width	Height
0	0.742188	0.517708	0.309375	0.735417
1	0.276562	0.529167	0.303125	0.725000
2	0.578125	0.547917	0.100000	0.520833
3	0.442969	0.576042	0.110937	0.510417


labelImg D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_1.png [2 / 1587]

D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_0.png
D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_1.png
D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_2.png
D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_3.png
D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_4.png
D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_5.png
D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_6.png
D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_7.png
D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_8.png
D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_9.png
D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_10.png
D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_11.png
D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_12.png
D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_13.png
D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_14.png
D:\Yolo\yolov5\new_data_0_images\train\images\Scenario_0_frame_15.png

Click & drag to move shape 'A'

X: 366; Y: 404

FIGURE 8. Operation interface of rectangular box label.

C. DATA ANNOTATION

People identification is a supervised learning task as discussed by [53]. Using the annotation tool LabelImg [54], we manually mark distinct individuals, encapsulating each person within a rectangular bounding box, illustrated in Figure 8. This tool produces annotations in the YOLO format, saving them as.txt files. Within these YOLO-format annotation files, each image's objects are described over several columns, where each column corresponds to a distinct object. Specifically, each line provides the subject ID, x-center, y-center, width, and height of the object. Where:

- *Subject ID:* It is an integer indicating the category index of the object (for example, 0 for "car", 1 for "person", etc.).

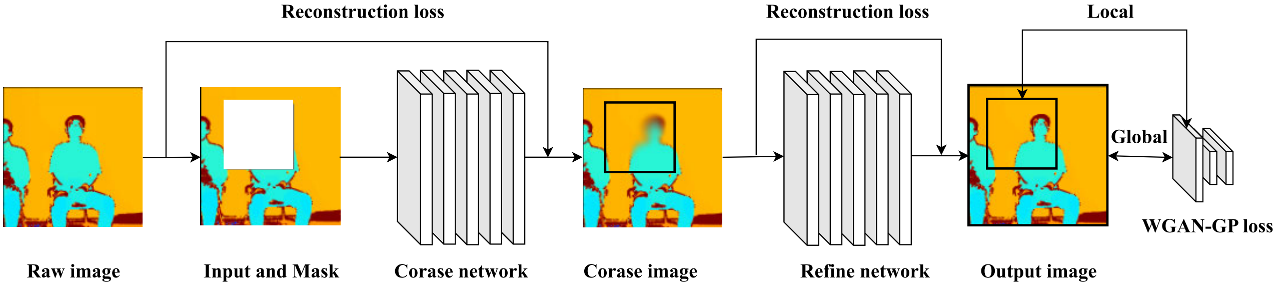


FIGURE 9. An overview of generative inpainting framework: Reconstruction loss is utilized to train the coarse network, refine network are trained by WGAN-GP global and local adversarial loss, as well as reconstruction loss.

TABLE 4. Annotation category information of YOLO.

Category	Subject ID
Subject A	0
Subject B	1
Subject C	2
Subject D	3

- *X center, Y center:* Represent the center coordinates of the object, normalized relative to the image's width and height. This means these values lie between 0 and 1.
- *Width, Height:* It is the width and height of the object, also normalized with respect to the width and height of the image.

A representative sample of our project's YOLO-format annotation can be found in TABLE 4. Detailed information regarding the categorization employed in our YOLO annotations throughout the experiment is presented in TABLE 5.

D. DATA AUGMENTATION

3D LiDAR devices primarily capture depth information of objects by gauging the distance and angle between the object and the device. However, there is an inherent limitation to this method: as the object's distance from the LiDAR increases, the accuracy of the information captured diminishes, often rendering the data sparse [55]. Consider a private car setting for instance: passengers in the rear seats are positioned farther from the 3D LiDAR compared to those in the front. Moreover, those in the back are frequently obscured by front-seat occupants, resulting in significant data occlusion. Our proposed solution to overcome this problem involves inpainting parts of the body that are occluded, thereby enhancing the process of identification.

In our study, we adopt the generative image inpainting method fortified with contextual attention. The architecture of the inpainting network can be visualized in Figure 9 and is characterized by a two-stage design. The initial part 'coarse network', focuses on generating a coarse prediction of the image. This is achieved using a reconstruction loss to guide its training. Following this, the 'fine network' steps in to refine the predictions. Unlike the coarse network, the fine network's training harnesses both the reconstruction loss and a Generative Adversarial Network (GAN) loss. Diverging from traditional generative inpainting techniques that leaned on Deep Convolutional GAN (DCGAN) [56] for adversarial supervise, our chosen algorithm turns to Wasserstein GAN Gradient Penalty (WGAN-GP) [57], [58]. Because when the performance gap between the generator and the discriminator is very large, the gradient disappearance problem occurs, causing training to stop. As such, WGAN maintains a balance between the generator and the discriminator by limiting the training of the discriminator to certain methods like weight clipping, thus helping to maintain training stability and continuous progress. Besides that, the traditional GAN loss function is usually based on the Jensen-Shannon scatter [59], which can lead to unstable training. To ensure both global and local consistency, we apply the WGAN-GP loss to the second-stage refinement network's global and local outputs. WGAN uses the Wasserstein distance [57] as the loss function, which is a measure of the difference between two probability distributions. The Wasserstein distance provides a much smoother gradient even when there is a large difference between the real data and the generated data. It leads to a more stable training process. Incorporated into the second phase of our network, the WGAN-GP loss combines the *l1* reconstruction with the *l1* distance metric, what is termed the Wasserstein-L distance $W(\mathbb{P}_r, \mathbb{P}_g)$. The model is constructed by applying Kantorovich-Rubinstein duality [60] and drawing on the improvement of Gulrajani et al. [58] who upgrade WGAN with a gradient penalty term, which is shown in Eq (5):

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} E_{(x,y)\sim\gamma}[\|x - y\|], \tag{5}$$

where $W(\mathbb{P}_r, \mathbb{P}_g)$ is the Wasserstein distance between two probability distributions $\mathbb{P}_r$ and $\mathbb{P}_g$, inf is the infimum, γ is the joint distribution between $\mathbb{P}_r$ and $\mathbb{P}_g$. $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ is the set of all possible joint distributions that couple $\mathbb{P}_r$ and $\mathbb{P}_g$. $E_{(x,y)\sim\gamma}$ is the expectation, which is over x and y randomly drawn from the joint distribution γ. Finally, $[\|x - y\|]$ is the Euclidean distance between x and y.

Contextual attention is another important part of the image reconstruction model, which is shown in Figure 10: First, a 3×3 patch is extracted from the background region, which is

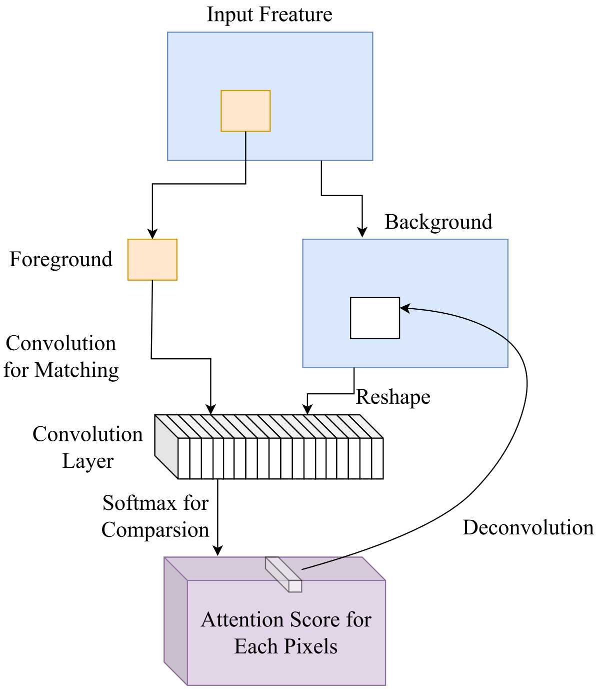


FIGURE 10. Illustration of the contextual attention layer.

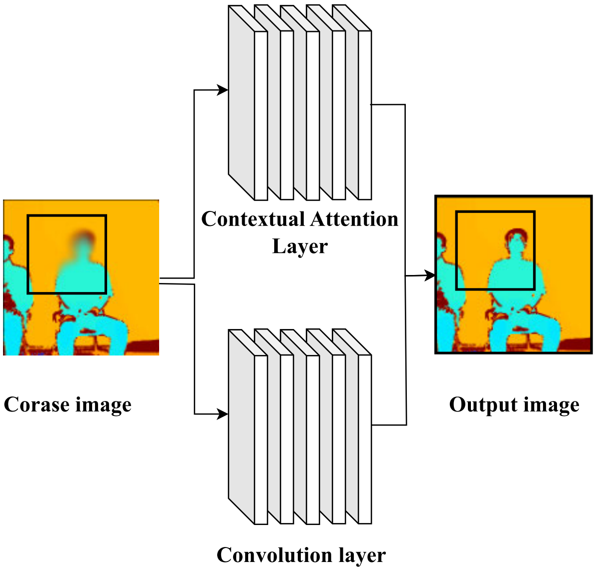


FIGURE 11. Illustration of detailed refine network.

used as a convolution kernel, in order to match the foreground (missing region) patch, a standardized inner product (residual selective similarity) is used to measure it, then Softmax is used to compute the weights for each background patch, and finally the best patch is selected and deconvolved to produce the foreground region. The overlapped pixels are averaged for the deconvolution process. To ensure greater image consistency, perceptual propagation is used. This technique involves applying an offset to the foreground region, aligned with a similar offset in the background. This is achieved using the unit matrix as a convolutional kernel. Contextual attention employs a two-step approach: first propagating left to right and then up to down. This dual-phase approach refines the attentional score. Notably, this method significantly improves the outcomes of restoration and introduces richer gradients during the training phase.

For the integration of the perceptual network into the second stage of the restoration network, a dual-branch architecture is devised within the second stage's repair structure. This architecture is illustrated in Figure 11. The lower branch is tasked with restoring the content of the missing areas using expansive convolution. In contrast, the upper branch is dedicated to extracting regions of interest from the background. In the final step, the outputs of both branches are merged using a concatenate function. This combined output is then passed through a decoder and processed with reverse convolution to produce the final outcome.

In our approach, we begin by viewing the subjects in the foreground (those seated in the front seat) as obstructions. The image regions corresponding to their bodies are pinpointed to create masks for individuals positioned in the back seats. Our method of choice for this task is the Canny algorithm [61]. For a start, we transform the image into grayscale, streamlining the subsequent enhancement stages. To augment the fidelity of the segmentation process, we implement thresholding [4]. This step elevates the image's contrast, paving the way for more precise edge detection. Upon successful edge identification, this data is harnessed to craft a segmentation mask for the image. This mask is meticulously layered onto the original image, ensuring the preservation of the initial pixel values in the delineated regions.

We fine-tuned the implementation of Yu et al. [45], [46] which is an open-source framework for generative image inpainting tasks. This is a popular approach based on deep generative models, adept at crafting unique image structures. For our training purposes, we sourced the ground truth data from scenario 3. This dataset comprises 7,000 images. To ensure our model's integrity and performance, we segregated an additional 1,000 images specifically for the validation phase. All our training exercises were conducted adhering to a specific environment configuration, which is detailed in Table 6. This table presents an exhaustive list of parameters, settings, and configurations used, ensuring reproducibility and a clear understanding of our training conditions. The trained model was provided with original images paired with their respective mask images. This setup was instrumental as it enabled our model to process these images with the primary objective of removing individuals seated in the front while simultaneously reconstructing those in the back seats. For a visual representation of what our model achieved, the examples are referred to in Figure 12.

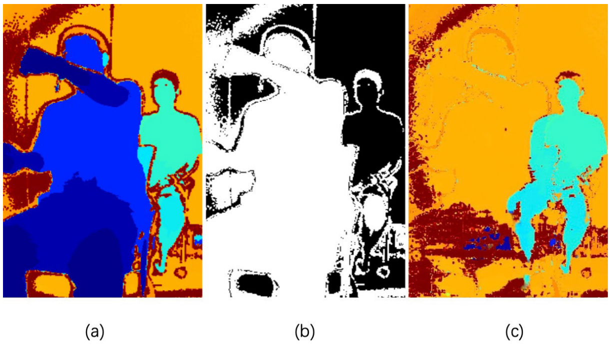


FIGURE 12. An example of people reconstruction: Image (a) is the original picture, image (b) is a mask segment extracted from people in the front, and image (c) is people in the back seat after reconstruction.

E. OCCLUSION RATE

Detection performance in private car surveillance is instrumental for ensuring passenger safety and comfort, especially for those seated in the back [62]. An essential metric for evaluating this is the occlusion rate, which quantifies how much of a backseat passenger is obscured from view. To accurately compute this rate, the study focuses on determining the ratio of the hidden portion of a passenger to their total visible area. Achieving this requires precise identification of the passenger's contours or boundaries within the frame, a task accomplished using image segmentation and edge detection techniques.

The methodology employed initiates with image segmentation via Gaussian blur [63], a technique that smoothens an image by reducing noise and minor details. By doing so, significant boundaries in the image become more pronounced. Following this, the Canny Edge Detection method [61], renowned for its efficiency in highlighting large intensity changes in images, is applied to ascertain these boundaries. Once these edges are defined, it becomes feasible to determine the area representing the passengers. The occlusion rate is then computed using a specific formula, Eq. (6), which compares the visible area to the total possible area, providing a percentage-based occlusion rate.

$$O_r(\%) = 1 - \frac{A_n}{A_t} \cdot 100, \tag{6}$$

where O_r refers to occlusion rate, A_n is the area of the not occluded part of the passenger, and A_t is the total area of the passenger.

V. EXPERIMENT

A. EXPERIMENTAL CONFIGURATION

The experiments' configurations are in TABLE 5. The implementation of YOLOv5 used in this work is that of Jocher et al. [8]

B. EVALUATION METRICS

Given that varying experiments yield different models, it is essential to establish a robust metric. This aids in the selection of the most effective models from the entirety of the experimental outcomes. In our experiment, we use the F1 score calculated according to Eqs. (7) to (9), and mean Average Precision (mAP) calculated at the Intersection over Union (IoU) threshold of 50% (mAP50) and 50%-95% (mAP50-95),

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{8}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{TP}{2 \cdot TP + FP + TN}. \tag{9}$$

Precision represents the proportion of true positive predictions among all positive predictions. *TP* (True Positives) denotes the number of positive instances correctly predicted as positive, while *FP* (False Positives) represents the number of negative instances incorrectly predicted as positive. A high precision indicates that most of the predicted positive instances are indeed positive. Recall also known as sensitivity or true positive rate, recall measures the proportion of actual positive instances that were correctly identified by the model. *FN* (False Negatives) denotes the number of positive instances incorrectly predicted as negative. A high recall means that the model correctly identifies most of the actual positive instances. The F1 Score is the harmonic mean of precision and recall. It provides a balanced measure between precision and recall, especially useful when there is a significant class imbalance. A high F1 Score indicates a well-performing model in terms of both precision and recall.

TABLE 5. Configuration of experimental environment.

Name	Parameter
GPU	Quadro RTX 5000, 16384 MB
System	Windows 11
Operating memory	16 GB
Environment configuration	Python-3.7.16 torch-1.13.1+cu117

TABLE 6. The hyperparameters used during the training.

Name	Value
Batch size	4
Epoch	30
Learning rate	0.01
Optimizer	SGA
Weights	yolov5s.pt

C. RESULTS AND DISCUSSIONS

The main training hyperparameters are shown in TABLE 5. After training, TABLE 7 summarizes the information including the F1 score, the value of mAP50, and mAP50-95 of each person. The valid is the number of validation datasets. The instances mean the number of people labeled in validation datasets.

TABLE 7. Results of people identification.

Class	Valid	Instances	F1	mAP50	mAP50-95
All	552	1671	0.665	0.753	0.626
A	552	552	0.517	0.632	0.48
B	552	453	0.711	0.823	0.684
C	552	115	0.506	0.642	0.566
D	552	551	0.844	0.915	0.774

In TABLE 7, we present an evaluation of our detection and identification models. The results obtained showcase impressive performance benchmarks. Specifically, our proposed methods registered an F1 score exceeding 60%. Additionally, the mAP50 score surpasses the 75%. This indicates the robustness of our approach in confidently identifying objects with a high degree of overlap. Furthermore, our model's ability to maintain a high score in the mAP50-95 range (surpassing 60%) signifies its consistency and accuracy across various IoU thresholds, illustrating its adaptability in different detection scenarios.

Figure 13 offers a more visual representation of our findings. In this figure, readers can witness direct instances of people detection and people identification, showcasing the practical application and effectiveness of our proposed models in simulation real-world scenarios. Meanwhile, Figure 14 delves deeper into the model's training outcomes. Here, we present the confusion matrix, which provides a granular view of the true positive, false positive, true negative, and false negative rates. According to Figure 14, subject D had the best prediction performance with a correct classification ratio of 0.90, meaning that 90% of the true subject D was correctly predicted to be D. Subject B had the next highest percentage of correct classifications at 0.76. Subject C has a correct classification ratio of 0.75. Subject A has a relatively low percentage of correct categorization of 0.56. The proportion of subject A misclassified as subject C was 0.18, which is a relatively high proportion of misclassification. It may be because Participants A and C are Asian males while Participant B is a white male and Participant D is an Asian female whose body shapes are relatively different between A and C. Therefore, subject A will easily be misclassified to be subject C. This matrix serves as a testament to the YOLOv5 model's proficiency in discerning and accurately identifying individuals in 3D LiDAR imagery.

It is also essential to highlight that the detection F1 scores and mAP values are not uniform across all individuals in our dataset. These values exhibit variations, primarily influenced by the occlusion levels associated with each individual. Such occlusions could arise from a myriad of scenarios, including overlapping individuals, obstructing objects, or even the angle of capture.

Taking into consideration the possibility of some individuals being obscured by others seated in front of them, we evaluated the F1 score, mAP50 values, and mAP50-95 values across varying occlusion rates in our experiments. Based on the dataset distribution and as calculated by Eq. (6), we observed that the occlusion rates for rear passengers fall within three ranges: below 60%, between 60% and 70%, and above 70%. The data distribution across these three intervals is fairly consistent. A comprehensive visualization of the data distribution can be found in Figure 15. We separately conducted F1 score, mAP50 values, and mAP50-95 values for each person in the left and right positions of the back row

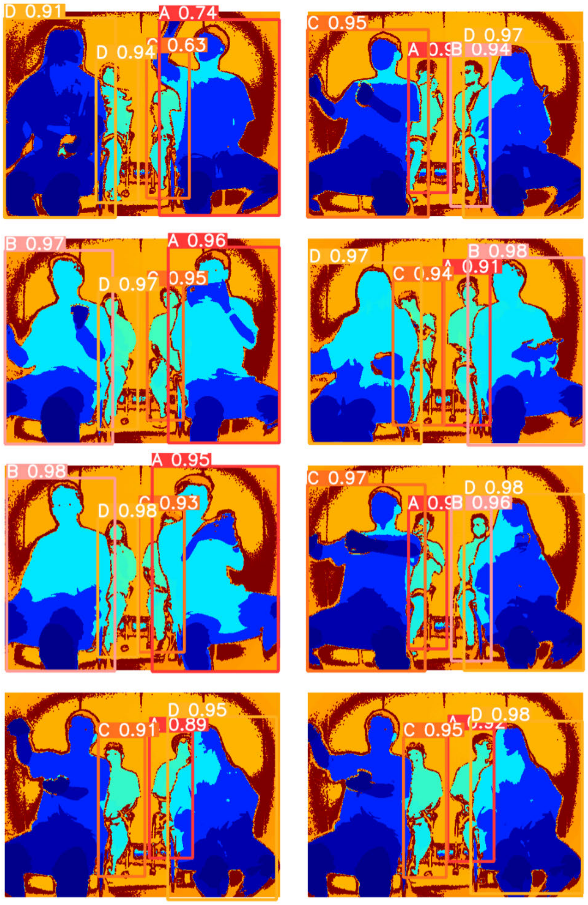
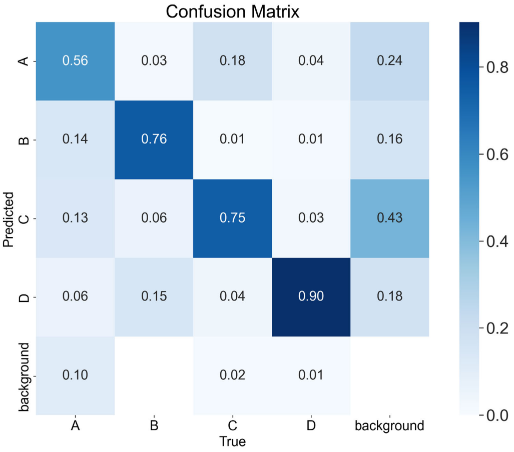


FIGURE 13. Some examples of people identification results.



Predicted \ True	A	B	C	D	background
A	0.56	0.03	0.18	0.04	0.24
B	0.14	0.76	0.01	0.01	0.16
C	0.13	0.06	0.75	0.03	0.43
D	0.06	0.15	0.04	0.90	0.18
background	0.10		0.02	0.01	

FIGURE 14. Confusion matrix of training results.

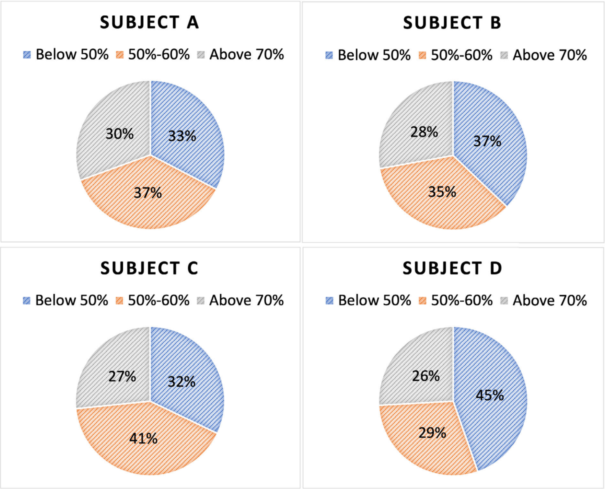


FIGURE 15. Data distribution of different occlusion rates.

with different occlusion rates, the specific data are recorded in TABLE 8. The 'OR' column means 'Occlusion Rate', the 'BL' means 'Back Left' and the 'BR' means 'Back Right'. The None means the quantities of data lay in this range are not enough for training so there is no data in this range.

As we can see in Figures 16 and 17, which is the illustration of the bar chart of the specific values of F1 score, mAP50 values, and mAP50-95 values at different occlusion rates for each person. In general, the performance metrics, particularly the F1 score, mAP50 values, and mAP50-95 values, tend to decrease as the occlusion rate increases. This observation is consistent across different individuals in the dataset, indicating that higher occlusion rates make it more challenging to achieve accurate detection and identification. It can explain subject D exhibits superior performance in people detection and identification, a phenomenon primarily attributable to the predominance of its occlusion rate being below 60%. This lower occlusion rate naturally facilitates improved detection and identification metrics. Additionally, a unique aspect to consider is that subject D is the only female participant in our experiment. Her distinct body shape, which stands out from her male counterparts, potentially augments the accuracy of people identification, serving as a unique and distinguishing feature for our proposed module.

In order to solve the problem that some individuals might be occluded by others seated in front of them that affects the training result of people identification. In our experiments, we use the YOLOv5 model to train the model on both pre and post-reconstruction images of participating people in the right and left back seats separately. The training environment is in TABLE 6. We then compare the F1 scores and mAP values of these instances. The results are presented in TABLE 10 and TABLE 10.

From TABLES 8 and 10, it is clear that the implementation of the generative image inpainting technique markedly enhances the performance of our people recognition model, we can see the bar chart before and after reconstruction in Figure 18. Such reconstruction not only elevates the F1 score but also bolsters the mAP50 values and mAP50-95 values metrics across all test subjects. This improvement can be attributed to the successful mitigation of visual obstructions that previously hindered clear views of individuals in the

TABLE 8. Training results from different occlusion rates.

Class	OR	Seat	F1	mAP50	mAP50-95
A	Below 60%	BL	0.816	0.91	0.465
	60%-70%		0.755	0.876	0.273
	Over 70%		0.661	0.511	0.353
	Below 60%	BR	0.664	0.857	0.655
	60%-70%		0.661	0.483	0.246
	Over 70%		0.587	0.448	0.267
B	Below 60%	BL	0.785	0.917	0.575
	60%-70%		0.616	0.762	0.602
	Over 70%		None	None	None
	Below 60%	BR	0.454	0.332	0.263
	60%-70%		0.417	0.324	0.236
	Over 70%		0.374	0.263	0.182
C	Below 60%	BL	0.659	0.75	0.461
	60%-70%		0.530	0.568	0.364
	Over 70%		0.313	0.464	0.173
	Below 60%	BR	0.903	0.995	0.625
	60%-70%		0.851	0.936	0.595
	Over 70%		None	None	None
D	Below 60%	BL	0.656	0.712	0.401
	60%-70%		None	None	None
	Over 70%		0.645	0.704	0.312
	Below 60%	BR	0.728	0.877	0.63
	60%-70%		None	None	None
	Over 70%		None	None	None

TABLE 9. Comparing the results before and after the reconstruction, people are in the left back seat.

Class	Instances	F1	mAP50	mAP50-95
Before Reconstruction				
A	70	0.707	0.787	0.501
B	50	0.866	0.961	0.810
C	50	0.774	0.788	0.519
D	50	0.601	0.619	0.265
Overall	220	0.737	0.788	0.523
After Reconstruction				
A	70	0.830	0.912	0.534
B	50	0.998	0.995	0.881
C	50	0.944	0.974	0.809
D	50	0.770	0.882	0.279
Overall	220	0.885	0.941	0.626

TABLE 10. Comparing the results before and after the reconstruction, people are in the right back seat.

Class	Instances	F1	mAP50	mAP50-95
Before Reconstruction				
A	70	0.670	0.569	0.220
B	50	0.488	0.379	0.248
C	50	0.894	0.919	0.601
D	50	0.567	0.647	0.432
Overall	220	0.655	0.628	0.375
After Reconstruction				
A	70	0.892	0.965	0.770
B	50	0.618	0.960	0.712
C	50	0.998	0.995	0.782
D	50	0.893	0.956	0.770
Overall	220	0.850	0.969	0.759

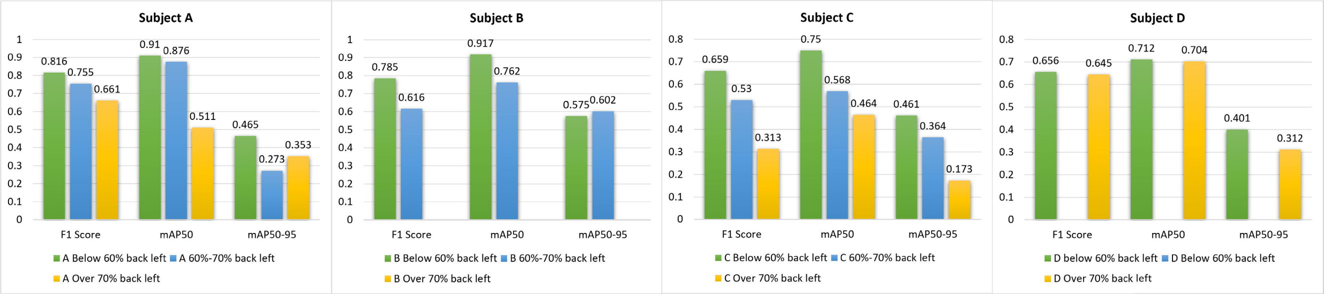


FIGURE 16. F1 scores, mAP50 values, and mAP50-95 values associated with different occlusion rates in the back left seats.

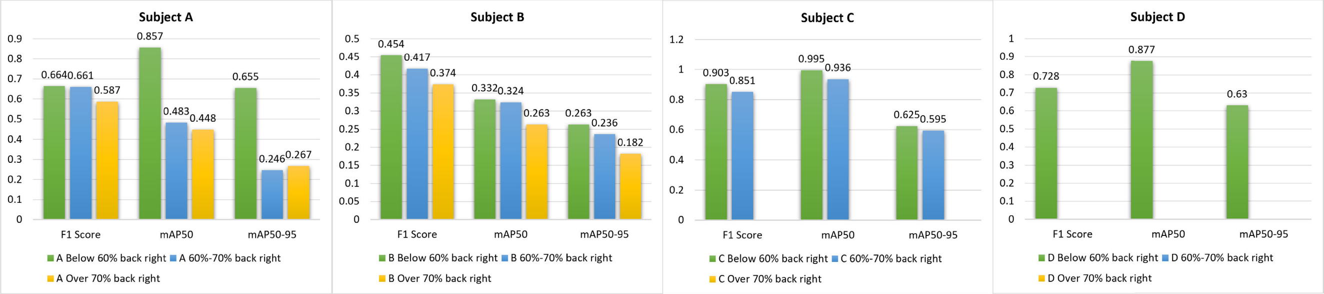


FIGURE 17. F1 scores, mAP50 values, and mAP50-95 values associated with different occlusion rates in the back right seats.

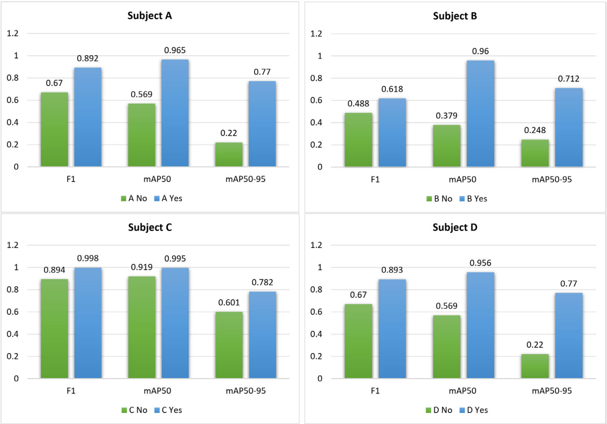


FIGURE 18. Comparison between before and after image reconstruction.

back seats. Moreover, any body segment earlier occluded by these impediments has been reconstructed, resulting in a more accurate and comprehensive body shape of the individuals.

Based on the results, the scores for the participants showed notable enhancements post-reconstruction. For subject A, the score increased from 0.707 to 0.830; for subject B, it increased from 0.866 to 0.998; for subject C, it increased from 0.774 to 0.944; and for subject D, it increased from 0.601 to 0.770. Similarly, when examining occupants in the vehicle's right back seat, the F1 scores experienced significant boosts post-reconstruction: subject A's score increased from 0.670 to 0.892, subject B's increased from 0.488 to 0.618, C's increased from 0.894 to 0.998, and subject D's increased from 0.567 to 0.893. The overall F1 score increased by 17.1%. About the mAP values, before reconstruction, subjects A, B, C, and D in the back left seat had mAP50 values of 0.787, 0.961, 0.788, and 0.619, respectively. Their corresponding mAP50-95 values were 0.501, 0.81, 0.519, and 0.265. After the reconstruction process, there was a noticeable enhancement in these metrics. Subject A's mAP50 increased to 0.912 with mAP50-95 0.534. Subject B's mAP50 slightly decreased to 0.995 but had an improved mAP50-95 of 0.881. Subjects C and D's mAP50 values increased to 0.974 and 0.882, respectively, with corresponding mAP50-95 improvements to 0.809 and 0.279. In the back right seat, Prior to reconstruction, subjects A, B, C, and D posted mAP50 scores of 0.569, 0.379, 0.919, and 0.647 respectively, and mAP50-95 values of 0.220, 0.248, 0.601, and 0.432. Post-reconstruction, all subjects witnessed improvements. Subject A's metrics rose to an mAP50 of 0.965 and an mAP50-95 of 0.77. Subject B's values surged to 0.96 and 0.712 for mAP50 and mAP50-95, respectively. Subject C, while having a marginal rise in mAP50 to 0.995, saw a slight decrease in mAP50-95 to 0.782. Subject D mirrored A's improvement, with mAP50 increasing to 0.956 and mAP50-95 to 0.77.

To the best of our knowledge, no research has been conducted with 3D LiDAR sensors in a manner identical to our experiments for people identification. While previous studies have utilized different sensor devices to collect 3D LiDAR images or methods to reconstruct 3D LiDAR images, we have provided a comprehensive comparison of overall system accuracy in TABLE 11. Yamada et al. [22]

TABLE 11. Performance comparison with state-of-the-art technologies.

Method	Sensor	Model	Dataset	Dataset Type	Accuracy
Yamada et al. [22]	Microsoft Kinect	CNN+LSTM	PCG [22]	Depth images	0.718
Jin et al. [64]	PrimeSense	IResNet100	UMIST [65]	RGB-D images	0.780
Jin et al. [64]	PrimeSense	IResNet34	UMIST [65]	RGB-D images	0.802
Ours	3D+RGB IP67 Kit	YOLOv5	Own dataset	Depth images	**0.807**

conducted an in-depth experiment on LiDAR-based gait analysis. Similarly, Jin et al. [64] employed a depth plus generative adversarial network approach to generate pseudo RGB-D images for identifying individuals. Because other state-of-the-art methods use depth images without occlusion, we compare their accuracy with the result of our reconstruction experiments. The result shows that Yamada et al. [22] used a Kinect sensor and a CNN+LSTM model on the point cloud gait (PCG) dataset [22], which is created by them, to achieve an accuracy of 0.718. Jin et al. [64] conducted two studies using the PrimeSense sensor with different IResNet models, achieving accuracies of 0.78 and 0.802 on the UMIST dataset [65], which is a face database consisting of 564 images of 20 individuals. Our method, employing a 3D+RGB IP67 kit and the YOLOv5 model on our own dataset, achieved the highest accuracy of 0.807.

These results demonstrate that our proposed model significantly enhances the identification and detection accuracy, especially for back seat passengers who were initially occluded by front seat individuals.

In the present study, we have primarily simulated scenarios within a private car environment, and it remains to be seen whether our findings would be replicated in actual private car settings. Additionally, the dataset used and the number of participants involved in our experiment are somewhat limited. The dataset we employed is relatively small and not evenly distributed, potentially impacting the robustness and reliability of our results. For future research, it is essential to expand the scope of data collection, ensuring a more comprehensive and diverse dataset. This would include involving a greater number of participants from different backgrounds to enhance the generalizability of our findings. Furthermore, conducting extensive experiments in real-world settings, particularly in actual private cars, will be crucial to validate and refine the effectiveness of our model. Such efforts are expected to provide more concrete insights and potentially lead to more robust and adaptable solutions for the challenges identified in our current research.

VI. ABLATION EXPERIMENT

To validate the effectiveness of our proposed method, we conducted ablation experiments on RGB datasets. These experiments were set up under the same conditions as described in Section IV. A. We established different lighting scenarios, as depicted in Figure 19 and Figure 20, to observe the differential outcomes yielded by our method in each distinct lighting environment. The result is shown in TABLE 12 and TABLE 13.

FIGURE 19. **Experiments in a lit environment.**

FIGURE 20. **Experiments in a dark environment.**

TABLE 12. **Results of people identification in RGB images in a lit environment.**

Class	Valid	Instances	F1	mAP50	mAP50-95
All	400	1597	0.937	0.941	0.782
A	400	400	0.839	0.846	0.712
B	400	399	0.990	0.995	0.828
C	400	399	0.940	0.936	0.786
D	400	399	0.979	0.99	0.804

TABLE 13. **Results of people identification in RGB images in a dark environment.**

Class	Valid	Instances	F1	mAP50	mAP50-95
All	400	1597	0.347	0.349	0.248
A	400	400	0.343	0.354	0.158
B	400	399	0.350	0.358	0.291
C	400	399	0.336	0.331	0.259
D	400	399	0.360	0.350	0.285

According to TABLE 13 and TABLE 14, the identification results are good in a well-lit environment. The overall F1 score is 0.937, the overall mAP50 value is 0.941, and the overall mAP50-95 value is 0.596. On the contrary, in a dark environment, the identification results are not

TABLE 14. Comparison of the results in RGB images in a different environment and in depth images.

Class	Instances	F1	mAP50	mAP50-95
RGB images in a lit environment				
Overall	1597	0.937	0.941	0.782
RGB images in a dark environment				
Overall	1597	0.347	0.349	0.248
Depth images				
Overall	1671	0.665	0.753	0.626

great. The overall F1 score is 0.347, the overall mAP50 value is 0.349, and the overall mAP50-95 value is 0.248. By comparing the training outcomes using RGB images and 3D LiDAR experiment, the results of RGB images in a lit environment perform better than the training outcomes of depth images. However, after image reconstruction, the results of depth images are much closer to the RGB outcomes in a lit environment, the overall results of the comparison are shown in TABLE 14. The results from depth images, demonstrate superior performance in dark environments. This indicates that our proposed method effectively addresses the issue of diminished detection effectiveness in a dark environment.

VII. CONCLUSION

In this paper, we showed the capabilities of the YOLOv5 model in detecting and identifying individuals within private car scenarios, using 3D LiDAR images. Traditionally, the utilization of 3D LiDAR often grappled with challenges, primarily its limited efficacy in gleaning comprehensive information about rear-seat passengers. The inherent design of vehicles, combined with the nature of the technology, rendered certain portions of these passengers obscured, thus presenting a clear challenge in achieving thorough recognition and identification.

To address this limitation, we integrated generative image inpainting technology into our methodology. This fusion of innovative reconstruction techniques with advanced detection models led to a significant enhancement in our results. Notably, metrics such as the F1 score, mAP50, and mAP50-95 values witnessed substantial improvements post-reconstruction, underscoring the transformative potential of combining these technologies for optimized people identification within private cars.

REFERENCES

[1] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Int. J. Comput. Vis.*, vol. 38, no. 1, pp. 15–33, 2000.

[2] D. M. Jang and M. Turk, "Car-Rec: A real time car recognition system," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2011, pp. 599–605.

[3] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[4] M. H. Chowdhury and W. D. Little, "Image thresholding techniques," in *Proc. IEEE Pacific Rim Commun. Comput. Signal Process.*, May 1995, pp. 585–589.

[5] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Aug. 2017.

[6] W. Shao, M. Bouazizi, and T. Ohtsuki, "3D-LiDAR-based people identification by generative image inpainting with YOLOv5 in private car scenario," in *Proc. IEEE ICIP*, Oct. 2024.

[7] X. Lu, J. Wu, X. Ren, B. Zhang, and Y. Li, "The study and application of the improved region growing algorithm for liver segmentation," *Optik*, vol. 125, no. 9, pp. 2142–2147, May 2014.

[8] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, Taoxie, J. Fang, Z. Yifu, C. Wong, D. Montes, Z. Wang, C. Fati, J. Nadar, V. Sonck, P. Skalski, A. Hogan, D. Nair, M. Strobel, and M. Jain, "Ultralytics/YOLOv5: V7.0—YOLOv5 SOTA realtime instance segmentation," Zenodo, Ultralytics, CA, USA, Tech. Rep., Nov. 2022, doi: 10.5281/zenodo.7347926.

[9] B. Douillard, J. Underwood, N. Kuntz, V. Vlaskine, A. Quadros, P. Morton, and A. Frenkel, "On the segmentation of 3D LiDAR point clouds," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 2798–2805.

[10] H. Guan, J. Li, S. Cao, and Y. Yu, "Use of mobile LiDAR in road information inventory: A review," *Int. J. Image Data Fusion*, vol. 7, no. 3, pp. 219–242, Jul. 2016.

[11] A. Asvadi, P. Girão, P. Peixoto, and U. Nunes, "3D object tracking using RGB and LiDAR data," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 1255–1260.

[12] X. Zhao, P. Sun, Z. Xu, H. Min, and H. Yu, "Fusion of 3D LiDAR and camera data for object detection in autonomous vehicle applications," *IEEE Sensors J.*, vol. 20, no. 9, pp. 4901–4913, May 2020.

[13] V. Bruce and A. Young, "Understanding face recognition," *British J. Psychol.*, vol. 77, no. 3, pp. 305–327, 1986.

[14] Y. Li and J. Ibanez-Guzman, "LiDAR for autonomous driving: The principles, challenges, and trends for automotive LiDAR and perception systems," *IEEE Signal Process. Mag.*, vol. 37, no. 4, pp. 50–61, Jul. 2020.

[15] J. Gómez, O. Aycard, and J. Baber, "Efficient detection and tracking of human using 3D LiDAR sensor," *Sensors*, vol. 23, no. 10, p. 4720, May 2023.

[16] M. Hasan, J. Hanawa, R. Goto, R. Suzuki, H. Fukuda, Y. Kuno, and Y. Kobayashi, "LiDAR-based detection, tracking, and property estimation: A contemporary review," *Neurocomputing*, vol. 506, pp. 393–405, Sep. 2022.

[17] W. Wang, X. Chang, J. Yang, and G. Xu, "LiDAR-based dense pedestrian detection and tracking," *Appl. Sci.*, vol. 12, no. 4, p. 1799, Feb. 2022.

[18] Z. Yan, T. Duckett, and N. Bellotto, "Online learning for 3D LiDAR-based human detection: Experimental analysis of point cloud clustering and classification methods," *Auto. Robots*, vol. 44, no. 2, pp. 147–164, Jan. 2020.

[19] Z. Yan, T. Duckett, and N. Bellotto, "Online learning for human classification in 3D LiDAR-based tracking," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 864–871.

[20] Z. Chen, T. Fan, X. Zhao, J. Liang, C. Shen, H. Chen, D. Manocha, J. Pan, and W. Zhang, "Autonomous social distancing in urban environments using a quadruped robot," *IEEE Access*, vol. 9, pp. 8392–8403, 2021.

[21] L. Li, X. Kong, X. Zhao, T. Huang, W. Li, F. Wen, H. Zhang, and Y. Liu, "RINet: Efficient 3D LiDAR-based place recognition using rotation invariant neural network," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 4321–4328, Apr. 2022.

[22] H. Yamada, J. Ahn, O. M. Mozos, Y. Iwashita, and R. Kurazume, "Gait-based person identification using 3D LiDAR and long short-term memory deep networks," *Adv. Robot.*, vol. 34, no. 18, pp. 1201–1211, Sep. 2020.

[23] J. Ahn, K. Nakashima, K. Yoshino, Y. Iwashita, and R. Kurazume, "2V-Gait: Gait recognition using 3D LiDAR robust to changes in walking direction and measurement distance," in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, Jan. 2022, pp. 602–607.

[24] M. Bouazizi, C. Ye, and T. Ohtsuki, "2-D LiDAR-based approach for activity identification and fall detection," *IEEE Internet Things J.*, vol. 9, no. 13, pp. 10872–10890, Jul. 2022.

[25] M. Bouazizi, C. Ye, and T. Ohtsuki, "Activity detection using 2D LiDAR for healthcare and monitoring," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6.

[26] M. Roth, D. Jargot, and D. M. Gavrila, "Deep end-to-end 3D person detection from camera and LiDAR," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 521–527.

[27] Y.-C. Fan, B.-T. Wu, C.-J. Huang, and Y.-H. Bai, "Environment detection of 3D LiDAR by using neural networks," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2019, pp. 1–2.

[28] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum PointNets for 3D object detection from RGB-D data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 918–927.

[29] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4490–4499.

[30] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[31] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[32] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[33] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[35] J. Kim, J. Kim, and J. Cho, "An advanced object classification strategy using YOLO through camera and LiDAR sensor fusion," in *Proc. 13th Int. Conf. Signal Process. Commun. Syst. (ICSPCS)*, 2019, pp. 1–5.

[36] J. Han, Y. Liao, J. Zhang, S. Wang, and S. Li, "Target fusion detection of LiDAR and camera based on the improved YOLO algorithm," *Mathematics*, vol. 6, no. 10, p. 213, Oct. 2018.

[37] K. S. Arikumar, A. Deepak Kumar, T. R. Gadekallu, S. B. Prathiba, and K. Tamilarasi, "Real-time 3D object detection and classification in autonomous driving environment using 3D LiDAR and camera sensors," *Electronics*, vol. 11, no. 24, p. 4203, Dec. 2022.

[38] M. Simon, S. Milz, K. Amende, and H.-M. Gross, "Complex-YOLO: Real-time 3D object detection on point clouds," 2018, *arXiv:1803.06199*.

[39] M. Simon, K. Amende, A. Kraus, J. Honer, T. Samann, H. Kaulbersch, S. Milz, and H. Michael Gross, "Complexer-YOLO: Real-time 3D object detection and tracking on semantic point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshop*, Jun. 2019, pp. 1–10.

[40] M. Simony, S. Milzy, K. Amendey, and H.-M. Gross, "Complex-YOLO: An euler-region-proposal for real-time 3D object detection on point clouds," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Sep. 2018, pp. 1–14.

[41] Y. Shao, Z. Sun, A. Tan, and T. Yan, "Efficient three-dimensional point cloud object detection based on improved complex-YOLO," *Frontiers Neurorobotics*, vol. 17, Feb. 2023, Art. no. 1092564.

[42] S. Y. Alaba and J. E. Ball, "A survey on deep-learning-based LiDAR 3D object detection for autonomous driving," *Sensors*, vol. 22, no. 24, p. 9577, Dec. 2022.

[43] H. Tian and L. Guo, "JYOLO: Joint point cloud for autonomous driving 3D object detection," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Oct. 2022, pp. 1–4.

[44] Q. Wu, X. Li, K. Wang, and H. Bilal, "Regional feature fusion for on-road detection of objects using camera and 3D-LiDAR in high-speed autonomous vehicles," *Soft Comput.*, vol. 27, no. 23, pp. 18195–18213, Dec. 2023.

[45] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5505–5514.

[46] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4470–4479.

[47] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," *Proc. Comput. Sci.*, vol. 199, pp. 1066–1073, Jan. 2022.

[48] L. O. Chua and T. Roska, "The CNN paradigm," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 40, no. 3, pp. 147–156, Mar. 1993.

[49] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[50] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[51] Y. Liu, Q. Wen, H. Chen, W. Liu, J. Qin, G. Han, and S. He, "Crowd counting via cross-stage refinement networks," *IEEE Trans. Image Process.*, vol. 29, pp. 6800–6812, 2020.

[52] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2018, pp. 8759–8768.

[53] M. Bäuml, M. Tapaswi, and R. Stiefelhagen, "Semi-supervised learning with constraints for person identification in multimedia data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3602–3609.

[54] D. Tzutalin, "LabelImg," *GitHub Repository*, vol. 6, 2015.

[55] G. L. Heritage and A. R. G. Large, "Principles of 3D laser scanning," in *Laser Scanning for the Environmental Sciences*. Oxford, U.K.: Wiley-Blackwell, 2009, pp. 21–34.

[56] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[57] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[58] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[59] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[60] S. T. Rachev, "Duality theorems for Kantorovich–Rubinstein and Wasserstein functionals," Ph.D. thesis, Instytut Matematyczny Polskiej Akademi Nauk, Warszawa, Poland, 1990.

[61] T.-S. Liu, R.-X. Liu, and S.-W. Pan, "Improved Canny algorithm for edge detection of core image," *Open Automat. Control Syst. J.*, vol. 6, no. 1, pp. 426–432, 2014.

[62] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Fast object detection with occlusions," in *Proc. 8th Eur. Conf. Comput. Vis. (ECCV)*, Prague, Czech Republic, 2004, pp. 402–413.

[63] R. A. Haddad and A. N. Akansu, "A class of fast Gaussian binomial filters for speech and image processing," *IEEE Trans. Signal Process.*, vol. 39, no. 3, pp. 723–727, Mar. 1991.

[64] B. Jin, L. Cruz, and N. Gonçalves, "Pseudo RGB-D face recognition," *IEEE Sensors J.*, vol. 22, no. 22, pp. 21780–21794, Nov. 2022.

[65] D. B. Graham and N. M. Allinson, "Characterising virtual eigensignatures for general purpose face recognition," in *Face Recognition: From Theory to Applications*. Berlin, Germany: Springer, 1998, pp. 446–456.

WEIRONG SHAO (Graduate Student Member, IEEE) received the B.E. degree in automation control from Beijing University of Chemical Technology, Beijing, China, in 2017, and the M.E. degree in electrical engineering from New York University, New York, NY, USA, in 2019. He is currently pursuing the Ph.D. degree in computer science with Keio University, Yokohama, Japan.

MONDHER BOUAZIZI (Member, IEEE) received the Bachelor of Engineering degree in communications from SUPCOM, Carthage University, Tunisia, in 2010, and the M.E. and Ph.D. degrees from Keio University, in 2017 and 2019, respectively. He was a Telecommunication Engineer (access network quality and optimization) for three years with Ooredoo Tunisia (Ex. Tunisiana). He is currently a Specially Appointed Senior Lecturer with the Faculty of Science and Technology, Keio University. He has published several journal and international conference papers. His research interests include machine learning, deep learning, data mining, sensors, and signal processing. He is a member of the Association for Computing Machinery (ACM) and the Institute of Electronics, Information and Communication Engineers (IEICE). He received the Telecommunications Advancement Foundation Student Award 2016, the IEEE/ACM ICSIM 2021 and 2022, the IEEE APPC 2021 Best Paper Award, and the A3 Workshop 2021 Best Presentation Award.

XIANG MENG (Graduate Student Member, IEEE) received the B.E. degree from the School of Automation Engineering, University of Electronic Science and Technology of China, in 2020, and the M.E. degree from Keio University, where she is currently pursuing the Ph.D. degree with the Graduate School of Science and Technology. Her research interests include fall detection based on 3D-LiDAR and deep learning.

TOMOAKI OHTSUKI (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees in Electrical Engineering from Keio University, Yokohama, Japan in 1990, 1992, and 1994, respectively. From 1994 to 1995 he was a Post Doctoral Fellow and a Visiting Researcher in Electrical Engineering at Keio University. From 1993 to 1995 he was a Special Researcher of Fellowships of the Japan Society for the Promotion of Science for Japanese Junior Scientists. From 1995 to 2005 he was with Science University of Tokyo. In 2005 he joined Keio University. He is now a Professor at Keio University. From 1998 to 1999 he was with the department of electrical engineering and computer sciences, University of California, Berkeley. He is engaged in research on wireless communications, optical communications, signal processing, and information theory. He is a recipient of the 1997 Inoue Research Award for Young Scientist, the 1997 Hiroshi Ando Memorial Young Engineering Award, Ericsson Young Scientist Award 2000, 2002 Funai Information and Science Award for Young Scientist, IEEE the 1st Asia-Pacific Young Researcher Award 2001, the 5th International Communication Foundation (ICF) Research Award, 2011 IEEE SPCE Outstanding Service Award, the 27th TELECOM System Technology Award, ETRI Journal's 2012 Best Reviewer Award, 9th International Conference on Communications and Networking in China 2014 (CHINACOM'14) Best Paper Award, 2020 Yagami Award, and The 26th Asia-Pacific Conference on Communications (APCC2021) Best Paper Award. He has published more than 266 journal papers and 490 international conference papers. He served as a Chair of IEEE Communications Society, Signal Processing for Communications and Electronics Technical Committee. He served as a technical editor of the IEEE Wireless Communications Magazine and an editor of Elsevier Physical Communications. He is now serving as an Area Editor of the IEEE Transactions on Vehicular Technology and an editor of the IEEE Communications Surveys and Tutorials. He is also serving as the IEEE Communications Society, Asia Pacific Board Director. He has served as general-co chair, symposium co-chair, and TPC co-chair of many conferences, including IEEE GLOBECOM 2008, SPC, IEEE ICC 2011, CTS, IEEE GLOBECOM 2012, SPC, IEEE ICC 2020, SPC, IEEE APWCS, IEEE SPAWC, and IEEE VTC. He gave tutorials and keynote speeches at many international conferences including IEEE VTC, IEEE PIMRC, IEEE WCNC, and so on. He was Vice President and President of the Communications Society of the IEICE, also he was a distinguished lecturer of the IEEE. He is a fellow of the IEICE and a member of the Engineering Academy of Japan.