Received 4 August 2023, accepted 24 September 2023, date of publication 28 September 2023, date of current version 3 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3320268

APPLIED RESEARCH

Time-of-Flight Depth Sensing for Automotive Safety and Smart Building Applications: The VIZTA Project

JASON RAMBACH, BRUNO MIRBACH, YURIY ANISIMOV, AND DIDIER STRICKER
German Research Center of Artificial Intelligence, Augmented Vision Department, DFKI GmbH, 67663 Kaiserslautern, Germany

Corresponding author: Jason Rambach (Jason.Rambach@dfki.de)

This work was supported in part by the Electronic Components and Systems for European Leadership (ECSEL) Joint Undertaking in collaboration with the European Union's H2020 Framework Program, in part by the Federal Ministry of Education and Research of the Federal Republic of Germany (BMBF) under Grant 16ESE0424, and in part by VIZTA under Grant GA826600.

ABSTRACT Time-of-Flight (ToF) can be an advantageous sensing modality for several indoor applications, used alone or in combination with other sensors such as RGB cameras. As part of the research project VIZTA (Vision, Identification, with Z-sensing Technologies and key Applications), we developed methods using Machine Learning algorithms with ToF depth measurements as inputs to address two key areas of applications, in-car cabin monitoring (person detection and segmentation) and smart building monitoring (person counting and anomaly detection). In this article, we discuss the entire research approach followed in VIZTA, from setting up the experimental environments for collecting data and creating the VIZTA public datasets, to developing Deep Learning algorithms tailored to ToF data, used either in 2D depth map or 3D point cloud format. We discuss the advantages and challenges of using ToF-data, as well as the lessons learned during the evaluation and benchmarking of our methods.

INDEX TERMS Building monitoring, deep learning, detection, in-car monitoring, machine learning, segmentation, time-of-flight.

I. INTRODUCTION

Time-of-Flight (ToF) sensors generate depth maps by measuring the time required for light to travel from the camera to a surface and back. Although there are some widely used sensors such as the Microsoft Kinect series using this technology, ToF sensing is not yet widely used in commercial and industrial applications. However, in the last few years, an increase in the use of ToF sensing is recognizable, e.g. through the tendency to incorporate ToF sensors in smartphone devices [1].

In fact, there are several advantages of ToF sensing compared to the commonly used RGB or monochrome cameras and other depth sensing techniques such as stereo. ToF can provide highly accurate depth at real-world scale even in feature-less environments, and is generally more resilient to illumination and color variations. The characteristics of ToF sensors make them especially attractive for indoor applications such as in-car or building interior monitoring. Another important factor are privacy considerations. Persons are much easier identifiable on RGB images than on depth maps, therefore depth sensing can be of interest for monitoring areas (person counting, anomaly detection etc.) without being invasive of peoples' privacy in public spaces.

Depth information in general, can be very beneficial as an input modality to a wide range of scene understanding problems of computer vision. Depth information is known to facilitate background removal and segmentation tasks as well as 3D understanding and reconstruction, for example in object pose estimation or SLAM [2], [3]. Another important point to consider is that realistic synthetic (3D rendered) depth data is much easier to generate than synthetic RGB data. This can be highly advantageous when large amounts of realistic data are needed to train neural networks. Nowadays, affordable sensors such as the Microsoft Kinect Azure [4] or

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

the Intel RealSense [5] are widely used in computer vision research.

It is in this vain that the VIZTA (**V**ision, **I**dentification, with **Z**-Sensing **T**echnologies and **A**pplications) EU Research project had been launched [6]. VIZTA aimed at developing innovative technologies in the field of optical sensors and laser sources for short to long-range 3D-imaging and to demonstrate their value in several key applications areas. Within VIZTA, 23 partners from Industry and Academia collaborated with the goal of developing new optical 3D Sensing devices (ToF, LiDAR) and use their outputs for developing improved algorithms in different areas of applications including automotive, security, smart buildings, mobile robotics for smart cities, and industrial maintenance. Our work in this project, described in this article, was to develop computer vision and machine learning algorithms using ToF cameras as input, for the areas of in-car cabin monitoring (person detection and segmentation, driver pose, attention and intention) and smart building monitoring (person detection, counting, anomaly detection). In contrast to the more widely used RGB-D variant, the goal was to explore the boundaries of machine learning models trained exclusively with the depth modality of ToF cameras, towards lower cost, privacy preserving systems.

The main contributions of this article is the summarizing of some of the first application-focused deep learning work done in VIZTA using exclusive ToF-depth input instead of using depth as an auxiliary modality to RGB. We provide end-to-end case study description for systems build for ToF starting from the experimental setup, data acquisition and annotation to the design of neural networks and their evaluation. Finally, our publically available benchmark ToF datasets available at https://vizta-tof.kl.dfki.de/ are unique in the field and currently being used to advance the topic of ToF-based perception. In this article, we summarize the most important results reported so far on these benchmarks as well as some highly interesting findings such as the ability to obtain results equivalent to RGB-D systems with a single ToF camera.

In the following we will first refer to the principles, history and future perspectives of ToF sensing in Section II. Subsequently, we describe our efforts in the VIZTA project, starting from the construction of experimental environments in Section III, the generation of the research datasets TICaM and TIMo in Section IV, the development of ToF depth image specific algorithms in Section V and finally the summary of all reported benchmark results on the VIZTA datasets to this day VII.

II. TIME-OF-FLIGHT CAMERA SYSTEMS

Time-of-Flight (ToF) sensors generate depth maps by measuring the time required for the light to travel from the camera to a surface and back. In this regard, the physical principle is the same as that of Lidar scanners. The difference is that the scene is not subsequently scanned to generate a depth map of the scene, but that the sensor performs the depth measurement parallel on a pixel matrix. ToF cameras have therefore special imaging sensors which correlate incoming light intensity with emitted light.

Therefore ToF cameras require a source which illuminates the scene with modulated infrared light. As the whole scene has to be simultaneously illuminated the light power density in the scene is low compared to that in a concentrated beam of a laser scanner. Therefore, ToF-cameras are only suited for short range application, with a typical working range of a few meters.

The range of ToF cameras corresponds to that of active stereo systems which rely on a pattern projector (see, e.g., the comparison of Kinect v1 and Kinect v2 [7]). Unlike those cameras, the illumination of the ToF cameras does not produce any particular pattern but is instead modulated in time. An advantage of ToF cameras is that they can be very compact, because there is no baseline required between the image sensor and the illumination.

The first ToF sensors that appeared around two decades ago had very low resolution of the order of 1k pixel because of the special pixel design including electronic circuits for the so-called demodulation of the incoming light [8]. However, over the last two decades new chip processes and pixel design allowed to increase the fill factor and demodulation contrast of the pixels, such that their size has been drastically reduced, thus increasing the sensor resolution (see [9] and references therein). In parallel, efficient infrared light sources, based on VCSEL technology, became available which are able to emit ultra-short light pulses or light continuously modulated with 100Mhz or more, which enhances drastically the precision of the depth sensing. State-of-the art ToF sensors have resolutions of up to 1Mpix and achieve frame rates of typically 30Hz up to 60Hz. The typical depth range is 5m. Depending on the sensor size, the pixel number and the FOV the range may be smaller or larger.

A breakthrough for ToF camera technology was the launch of the Kinect v2 as a component of the Xbox One by Microsoft in 2013, which replaced the structured light based first version [10]. Meanwhile the third device generation, the Microsoft Kinect Azure with further enhanced ToF sensor [4], [7] is available and already widely used in the research community.

The introduction of the Kinect v2 to the consumer market was the first application for a mass market of the ToF camera technology. Several other application fields for ToF sensors have been explored since then. Existing applications are:

- Building management: Monitoring system based on ToF sensors have been around for about ten years [11], which realize basic building management functions such as people counting or single access control.
- Automotive in cabin monitoring: As an option for luxury cars, BMW (2017) [12] and Daimler (2020) [13] offer ToF cameras integrated in the overhead module of the car which and can recognize hand, respectively arm gestures.
- Industry automation: Several camera manufacturers offer ToF camera systems for industry applications based on latest sensor technologies which offer resolutions up to 0.3Mpix, framerates of up to 60Hz, and a range of up to 10m [14], [15]

These are applications with a rather low market volume. However, the tendency to incorporate ToF sensors in the smart phone devices, has in recent years given a new momentum to the further development of the ToF technology, mainly with focus on energy- and size optimized sensors with high resolution [9].

From this progress in the sensor technology all application areas can potentially profit. In particular if the sensor technology is combined with state-of-the-art deep-learning algorithms as was done in the VIZTA project, novel functionalities can be realized.

III. EXPERIMENTAL ENVIRONMENTS

Due to the absence of existing ToF datasets corresponding to the planned tasks in the VIZTA project, it was necessary to built dedicated environments for data collection and testing of the developed solutions. Therefore, two experimental setups were created, corresponding to the in-car scenario and the building monitoring scenario respectively.

A. DRIVING SIMULATOR - AUTOMOTIVE IN-CABIN ENVIRONMENT

The in-cabin test platform [16] is based on a driving simulator consisting of a realistic in-cabin mock-up and a wide-angle projection system for a realistic driving experience. The test platform has been equipped with a wide-angle RGB-D camera system (Microsoft Kinect Azure [4]) for monitoring the entire interior of the vehicle mock-up from a position corresponding to the overhead module of a real car. In addition, an optical ground truth reference sensor system [17] that allows tracking and recording the occupant's body movements synchronously with the 2D and 3D video streams of the camera. Moreover, the precise positioning of the front seats can be controlled and registered via a CAN-interface. The OpenDS driving simulator [18] software is utilized with three projectors covering almost the driver's entire field of vision as can be seen in Figure 1.

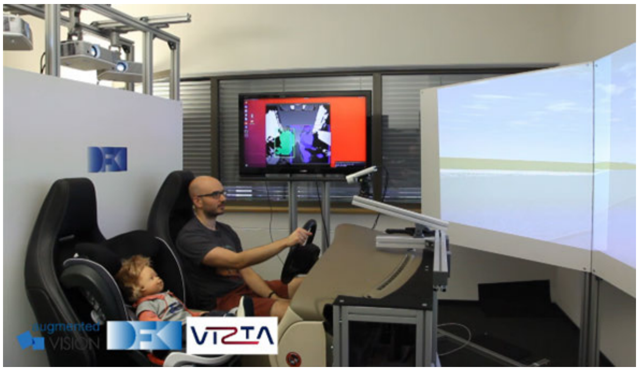


FIGURE 1. VIZTA in-cabin driving simulator experimental environment.

B. SMART BUILDING SETUP

Similarly to the in-car scenario, a dedicated setup was created for the building monitoring scenarios. The recording platform was equipped with a Kinect Azure RGB-D camera as well. Two separate views were designed. First, a top down view (bird's eye) with the camera mounted close to the ceiling, suitable for access control and person detection operations. Secondly, a tilted camera view, able to monitor a larger space, making it more suitable for behavior analysis operations such as anomaly detection. Depth maps from both views are shown in Figure 2.

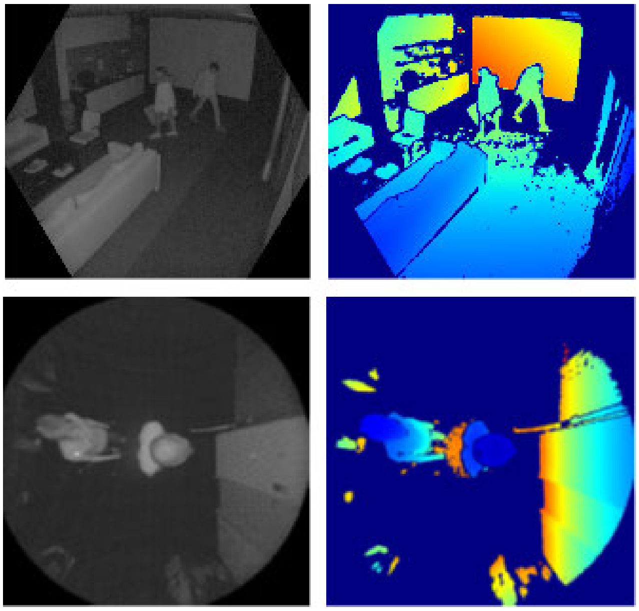


FIGURE 2. Top: IR images and depth maps from tilted camera view in building setup. Bottom: IR images and depth maps from the bird's eye camera-view.

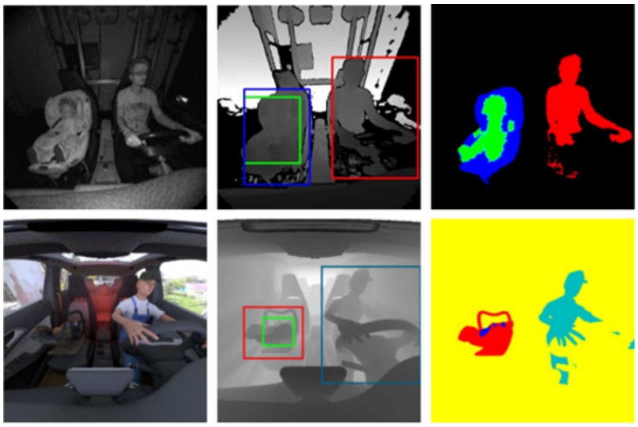


FIGURE 3. Sample data from the TICaM dataset. Top: Real data - IR image, Depth map with box annotation, Segmentation Masks. Bottom: Synthetic data - IR simulation, Depth map, Segmentation Mask.

IV. DATASETS

Due to the absence of any publicly available datasets of ToF camera images covering our target application areas of in-car cabin monitoring and smart building person and anomaly detection, we utilized our experimental setup environments described above in Section III in order to record two new

TABLE 1. Overview of main characteristics of all VIZTA ToF Datasets.

VIZTA ToF Datasets Overview			
-	**TICaM [19]**	**TIMo Anomaly [20]**	**TIMo Person detection [20]**
# Frames	$118K$	$612K$	$23.6K$
Camera View	Overhead module	Indoor top-down Indoor tilted	Indoor top-down
Modalities	RGB,Depth,IR	Depth,IR	Depth,IR
Labels	2D+3D Bounding Box ($10K$) 3D Segmentation Masks ($10K$) Per-frame Activity ($118K$) 2D Pose Keypoints ($3.3K$)	Per-frame Anomaly	2D+3D Bounding Box ($23.6K$) 2D Segmentation Masks ($23.6K$)
Classes	$16(6)$ Object, 19 Activity	2 (Normal - Anomalous)	1 (Person)
Data-Type	Real($115K$), Synthetic($3.3K$)	Real	Real

ToF datasets. The datasets TICaM (in-car) and TIMo (Indoor monitoring) were published in [19] and [20] respectively, and were made publicly available for research purposes through our website https://vizta-tof.kl.dfki.de/.

Both datasets use the Kinect Azure RGB-D camera. The annotation of the data was done manually, assisted by some semi-automatic functionalities of our annotation tool, such as the transfer of bounding box proposals between subsequent frames [21]. Table 1 summarizes the attributes of the two datasets.

A. TICAM - IN-CABIN DATASET

Due to the general interest in the in-car monitoring task, there is a significant number of publicly available research datasets in this domain [22], [23], [24], [25], [26], [27], however most of them focus on the task of activity recognition or head-pose estimation (no bounding box and segmentation labels), while their field of view only covers part of the car cabin, i.e. only the driver is monitored [22], [23], [24], [25], [27]. In addition, only [22] and [25] provide depth data but not from a ToF camera. The most relevant dataset to our work, in terms of specifications is SVIRO [26], however this dataset provides only synthetic data of cars' back seats.

Our dataset, TICaM [19], is the first dataset providing wide-angle ToF images with a FoV covering the entire car-cabin in three modalities (ToF Depth, ToF IR, RGB) and different occupancy scenarios (driver + passenger/object/child seat). A total of $118K$ frames are provided with per-frame person activity annotations, while $10K$ frames ($6.7K$ real, $3.3K$ synthetic) are annotated with object bounding boxes and segmentation masks. The dataset provides multi-level classification granularity, with the most commonly evaluated class-set being *(Person, Object, Child, Infant, Forward-facing child seat, Rearward-facing infant seat)*. The combination of both synthetic and real images in the same dataset is another unique feature of the TICaM dataset, enabling synthetic model training to real data inference transfer learning evaluation and domain adaptation strategy development. Sample data from the TICaM dataset are shown in Figure 3.

B. TIMO - INDOOR MONITORING DATASET

Same as in the case of in-cabin monitoring, we recorded and annotated our own dataset, TIMo [20] for our indoor monitoring cases. The dataset consists of two sub-parts:

- The **TIMo Anomaly Detection** dataset is designed for unsupervised anomaly detection, consisting of 1588 Sequences with a total of $612K$ ToF depth and corresponding IR frames from a top-down and a tilted camera view (See also Figure 2). The training set contains only normal data ($365K$ frames) while the test data contains $170K$ normal and $75K$ anomalous frames with frame-level anomaly annotations. All anomalies were acted and belong to three main categories, namely *Medical Emergencies, Violent Behavior and Left-behind Objects*. Selected IR frames from two sequences are shown in Figure 4.
- The **TIMo Person Detection** dataset addresses the needs of applications such as people counting or access control, and therefore provides ToF depth and IR data from a top-down camera view only, from 2 different locations. For supporting the generality of developed machine learning solutions, the height of the camera is varied between 2.25 and 2.75 meters. In total, more that $22K$ frames ($14K$ training, $8.6K$ testing) are provided, annotated with bounding boxes and segmentation masks for single-class classification of persons.

The TIMo datasets are also unique when compared to existing datasets in similar domains. Scene anomaly detection datasets are typically recorded outdoors [28] and often do not provide depth camera information [29], while some provide only action class annotations instead of anomaly information [30], [31].

V. ALGORITHM DEVELOPMENT

A. IN-CABIN DETECTION AND SEGMENTATION

In a number of research studies, we have investigated algorithms for occupant detection and segmentation based on ToF data, with focus on several questions, which are:

1) How to adapt network architectures originally developed for RGB images to depth data?
2) Which network architecture is most promising in terms of performance and real-time capability?
3) Which training strategy allows the optimal use of synthetic training data for adaptation to real evaluation data for detectors?

The most promising approach to achieve a real-time capable person and object detector is to adapt a state-of-the art 2D convolutional neural network based detector to depth data.

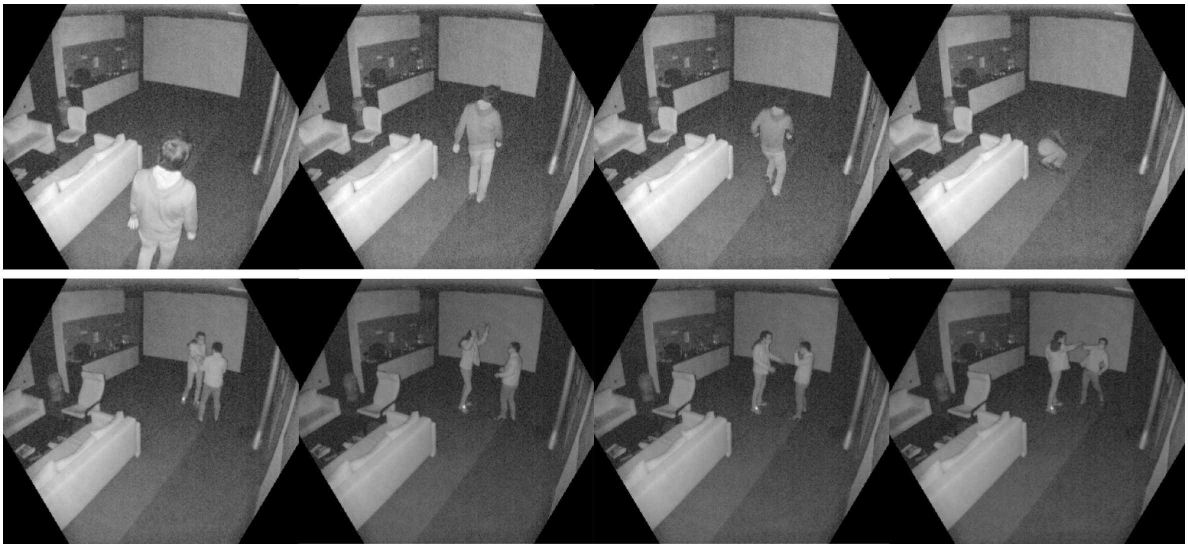


FIGURE 4. Example anomaly sequences from the TIMo Anomaly dataset, ToF IR modality. Top row: Medical emergency case (Fainting person). Bottom row: Aggressive beahviour case, arguing and fighting.

These detector approaches can be broadly divided into two categories based on their architecture. So-called two-stage detectors employ a 'propose regions first, then detect' approach, where several regions are first selected from the image based on their likelihood to contain an object. These regions are used to pool features within them and then to compute the location of the object and the probability that they belong to a certain class. Architectures like Faster R-CNN and its predecessors R-CNN, and Fast R-CNN fall in this category (see [32] and references therein).

Several state-of-the art instance segmentation methods are built upon these two-stage detectors. Mask R-CNN is, e.g. an extension of Faster R-CNN, which incorporates a parallel mask prediction branch [33].

The second class of object detectors are one-stage detectors, which treat object detection as a problem of regression. The objective is to directly arrive at the bounding box coordinates and the class probabilities of the objects from the feature maps extracted from the entire image. In this way, single-stage detectors such as YOLO, YOLACT and SSD have a significant higher detection speed with comparable accuracy compared to two-stage detectors (see [34]). The advantage of YOLACT compared to the other one-stage detectors is that it offers also an instance segmentation in addition to the detection. Because of its parallel architecture, the segmentation adds only small computational overhead to the detection. Thus, the YOLACT architecture, published in 2019 after the start of the VIZTA project, has become during the project the most promising candidate for a real-time in-cabin person and object detection and segmentation system.

The most straight-forward way to apply a 2D neural network originally developed for RBG images is to scale the depth image to the appropriate input value range and to replicate the values in all 3 channels. In the in-cabin algorithm algorithm depth data have been clipped to a relevant range which was chosen to be $[0, 2.55m]$ and then normalized to the input range of the model to maintain maximal precision. Alternative representation of depth data as surface normals or HHA representation may also be considered. In any case, one needs to be aware that the initial weights from a model pre-trained on RGB data may not be optimal for training the detector due to the different characteristics to depth data. A training from scratch on annotated depth data is therefore recommended.

Since the collection and labelling of training and validation data is a tedious and error-prone task, the question arises to what extent real data can be substituted by synthetic data. This idea becomes particularly relevant for the serial development of automotive in-cabin sensing systems where algorithms must be rapidly adapted to different vehicle models and variants. However, as long as the synthetic data do not reproduce all specific artefacts of real depth data caused by both the sensor and the scene, a network purely trained on synthetic data will be challenged to generalize well on real data.

In our work [35], we have investigated different strategies to incorporate synthetic data in the network training and explored an adversarial training based framework for adapting depth images from synthetic to real domain in an unsupervised manner. Synthetic training data were taken from the SVIRO dataset [26], while real test data and additional training data were taken from a subset of the later published TICaM dataset. Trained model architectures were Faster R-CNN and Mask R-CNN for detection and instance segmentation. Several baseline results were presented. The main observations and conclusions are:

- As expected, a network trained exclusively on synthetic depth data does not generalize well on real data. This negative impact of the domain shift between synthetic to real data is thereby larger on the segmentation precision than on the detection precision.

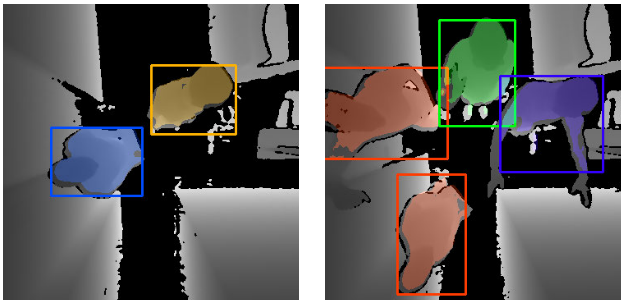


FIGURE 5. Examples of person detection from top-down perspective.

- A fine-tuning training strategy, in which the network is trained on combination of synthetic and real data, yields for both detection and segmentation a performance superior to a network trained on real data only.
- A refinement of synthetic data with a generative network for domain adaptation can substitute real data in training of a detector. However, an accurate segmentation still requires a fine-tuning with real data.

Therefore this study has clearly indicated the need for further research in the domain adaptation of depth data with the goal to minimize detail loss in real data which seems crucial for segmentation.

B. SMART BUILDING - PERSON DETECTION AND SEGMENTATION

In a smart building monitoring scenario, a vision-based system can have multiple uses with some of the most common applications including person counting on area entrances for crowd control or access control systems for tailgating avoidance. Detection of persons on camera images from a top-down perspective is typically the first computational step of algorithms developed for such building monitoring applications.

The use of ToF cameras, offers again some significant advantages compared to RGB cameras. Privacy preservation is enhanced since persons can be detected on depth images based on shape but are not easily identified (recognition of person identity). Furthermore, depth information is very valuable for person detection since it enables a natural background separation, and allows to accurately track the position of a person in a room. Finally, in indoor scenarios, ToF-based systems are more resilient to lighting variations being active light sensors and can therefore function well in very low-light conditions.

For the specific problem of person detection from ToF images, no prior work to ours was published. Our implementation of the person detector was based on an adaptation of the RGB-based YOLACT neural network [34] detector. Using the original ResNet101 backbone led to overfitting and false positives in the TIMo dataset, therefore it was replaced with a smaller and easier to train ResNet18 backbone. Examples from the algorithm output are shown in Figure 5.

C. SMART BUILDING - ANOMALY DETECTION

Anomaly detection at a per-frame level aims to mark specific camera frames of a video sequence that may contain anomalous events, such as criminal activity (violence, forgotten objects) or medical emergency situations. In building monitoring, it can serve in surveillance systems as an indication of a possible situation requiring intervention measures.

Anomaly detection is a challenging real-life application for machine learning since the search space is not deterministically defined, meaning that it is not possible to create supervision data for all possible target cases. Therefore, unsupervised learning is realistically the most promising approach for anomaly detection. Indeed, most existing techniques exploit the latent space representations learned exclusively from normal data. These learned representations can be used for reconstructing inputs with an autoencoder or for predicting the next frame in a sequence. It is then expected that when such trained networks are presented with abnormal data, they produce reconstructions or predictions of lower accuracy, which can then serve as an indication of an anomalous (out of distribution) incident in the video sequence [36].

The use of ToF depth as the sole modality for this application had not been investigated before our work in [37], even though there are considerable advantages. Most importantly, ToF depth offers privacy preservation, as discussed beforehand, while it is also more robust to lighting variations. Moreover, a ToF camera can be more compact, less costly and less energy-consuming than RGB-D variants.

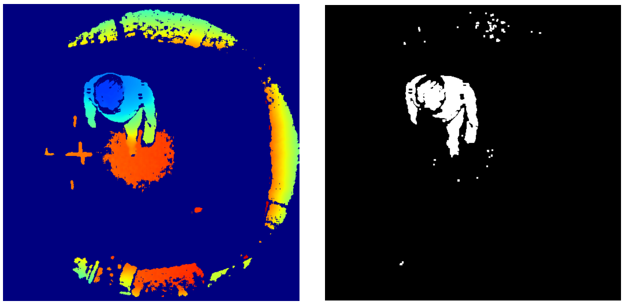


FIGURE 6. Depth map of the TIMo anomaly dataset (left) and corresponding extracted foreground (right).

In practice, surveillance systems of this type in buildings are typically set-up statically. This means that the scene is clearly separated into a static or slowly changing background and a dynamic foreground where the interaction between persons or between persons and objects happen. The targeted anomalies are dynamic events happening almost exclusively in the foreground.

Depth data offer another significant advantage here, namely that background-foreground segmentation is very simply possible by inspection of the depth data compared to an empty background image [38] (see Figure 6). Therefore, since anomalies happen in the foreground, our initial notion was to perform both the training and the inference for our anomaly detection algorithms on foreground images. However, this was proven to be ineffective, possibly due to loss of some interaction information or due to having too few depth values overall in the foreground images.

Therefore, in [37], we suggested to use a modified loss function that applies a weighting of the pixel reconstruction losses of autoencoders depending on whether they belong to the foreground or background. An example of the effect of applying this W-MSE (Weighted Mean Square Error) loss in reconstruction is shown in Figure 7. In this manner, the foreground reconstruction where most of the anomaly activity can happen is given priority, while the background is not completely ignored.

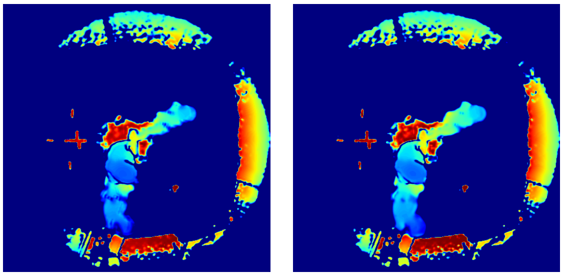


FIGURE 7. Autoencoder reconstruction when a normal Mean-Square Error (MSE) is used in training (left) vs. the Weighted Foreground MSE (right). W-MSE reduces slightly the reconstruction quality of the background (scene) while improving the person reconstruction.

VI. NEURAL NETWORK OPTIMIZATION FOR EMBEDDED SYSTEMS

The detection networks for the studies on person detection in buildings [19], [35], [39] and in-cabin occupant detection [20] were implemented in Python using the PyTorch framework with the possibility of further network conversion to ONNX format with opset 11. The training was done on an Ubuntu 18.04 × 64 PC with 4 Nvidia GTX 1080Ti Graphics Processing Units (GPUs). The inference evaluation on an embedded platform was performed on the Jetson AGX Xavier platform [40].

The detection network based on YOLACT network showed compared to two stage detectors, like Faster R-CNN, not only a higher detection accuracy, but also a faster inference speed. This algorithm runs at 0.03sec per frame on an Nvidia GTX 1080Ti GPU, making it suitable to process a depth camera stream captured at 30fps.

However to run the algorithm on an embedded GPU, like the Nvidia Jetson platform, the network size was reduced and a target-platform specific optimizations was performed.

To reduce the network size, the original ResNet101 feature extraction backbone network was replaced by ResNet18 which has a smaller number of layers. In this way, the network size has been reduced from 49.5 M parameters to 21 M. The reduced network size affects also the runtime, which was thereby reduced from 0.03s/frame to 0.025s/frame.

In addition, the smaller number of training parameters reduces the risk of overfitting. The smaller number of layers also allows to train the network with a larger batch size. The features from pre-trained ResNet on ImageNet are not used anymore, since their usage with depth data is not optimal. Instead, the weights were initialized randomly. According to our test configuration, random weight initialization converges to a lower number of false positive detections.

For deploying the detection algorithm on the embedded target platform NVIDIA Jetson, the trained model is first converted to the conventional ONNX format, which represents the network structure and contains the trained weights. For its inference, the model is passed through a TensorRT-based pipeline, which uses the platform-specific optimization for the generation of faster models with minimal accuracy drop. Based on the model definition the framework performs weights quantization to lower bit resolution and replace particular network function calls to the best ones for the desired GPU. It positively affects the running time of the YOLACT algorithm on embedded hardware, increasing the runtime performance from 11 to 30 fps.

VII. EXPERIMENTAL EVALUATION

Although our two introduced benchmark datasets TICaM [19] and TIMo [20] were only recently introduced, several research publications have already used them for evaluation and reported quantitative evaluation scores on them. In the following, we attempt to summarize and compare the heterogeneous results from all these publications as well as possible.

A. TICAM BENCHMARK

The TICaM dataset [19] was used in several works, with diverse research questions addressed. An overview of all reported results on the dataset is given in Table 2. Initially, in the TICaM publication [19], several baselines on the target tasks of object detection and segmentation using only the depth-map modality were given. The best results were achieved with an approach based on the YOLACT model [34], adapted to depth as discussed in Section V-A. As in the previous study [35], a very large domain gap can be seen when models are trained exlusively on synthetic data and applied on real data. When training on the real set or a combination of synthetic+real fine tuning the results were very promising taking into account the exclusive use of depth information without RGB or IR.

The work presented in [39], investigated the possibility of creating partial-view 3D point clouds out of depth images of TICaM, and performing segmentation using point cloud based methods such as PointNet [43], before mapping back to the depth image. The comparison of point cloud based approaches against depth-map based methods, showed that each have their own strengths and weaknesses, with PointNet++ achieving the highest mIoU value on TICaM so far.

The work of [45] argued that the use of multi-task networks with a shared backbone and branches for segmentation, detection and pose estimation is advantageous in the in-car monitoring task. DaCruz et al. [47] addressed the issue of domain adaptation from synthetic to real data. They suggested an auto-encoder based method that learns to reconstructs the input as support to the classification task. In their experiments, they achieved an impressive adaptation result in terms of detection mean Average Precision (mAP) when training on the purely synthetic SVIRO [26] dataset of back-seat views

TABLE 2. Summary of reported experimental results on VIZTA TICaM dataset for in-car cabin monitoring. † [45] uses a custom training/validation split on TICaM.

TICaM Dataset - In-car detection and segmentation							
Method	**Base/Backbone**	**Modalities**	**Training data**	**Classes**	**B.box mAP**	**Mask mAP**	**mIoU**
Katrolia et al. 2021 (TICaM) [19]	YOLACT [34] (ResNet101)	Depth	Real (TICaM)	6 (TICaM)	91.11	85.78	-
			Synthetic (TICaM)		27.67	18.99	-
			Real + Synthetic (TICaM)		87.12	86.25	-
Depth vs. Point cloud 2021 [39]	DeepLab v3 [41]	Depth	Real (TICaM)	6 (TICaM)	-	94.60	85.35
	FCN [42]				-	96.74	81.49
	PointNet [43]	Point Cloud			-	84.16	75.07
	PointNet++ [44]				-	91.19	87.32
Ebert et al. 2022 [45]	CSPNet [46]	IR	Synthetic (SVIRO [26]) + Real (TICaM)†	5 (SVIRO [26])	$49.40^{\dagger}$	-	$87.20^{\dagger}$
DaCruz et al. 2022 [47]	VGG-11 [48]	IR	Synthetic (SVIRO [26])	5 (SVIRO [26])	81.00	-	-
	ResNet-50 [49]				83.70	-	-
	DenseNet-121 [50]				79.30	-	-
Sharma et al. 2023 [51]	ShapeConv [52]	RGB-Depth	Real + Synthetic (TICaM)	6 (TICaM)	-	-	77.39
	ShapeConv [52]	IR-Depth			-	-	74.61
	MTL-DA ShapeConv [51]	IR-Depth			-	-	79.73

TABLE 3. Summary of reported experimental results on VIZTA TIMo dataset for the task of anomaly detection. * Modified CAE architecture compared to [36], more details in [20].‡ A Vision transformer autoencoder inspired by [54]. Implementation details in [37]. P- denotes prediction of next frame as anomaly detection approach, R- denotes reconstruction of current frame for anomaly detection.

TIMo Anomaly Dataset - Top-down view data				
Method	**Base/Backbone**	**Modalities**	**Training data**	**AUROC**
TIMo Dataset 2022 [20]	CAE* [36]	Depth	TIMo training set (normal data only)	56.4
	ConvLSTM [28]			62.2
Schneider et al. 2022 [37]	R-CAE* [36]	Depth	TIMo training set (normal data only)	**73.2**
	P-CAE* [36]			67.8
	R-ViT-AE‡ [54]			65.3
	P-ViT-AE‡ [54]			63.0
	P-ConvLSTM [28]			65.6
	R-CAE* [36]	IR		60.2
	P-CAE* [36]			57.7
	R-ViT-AE‡ [54]			57.7
	P-ViT-AE‡ [54]			57.1
	P-ConvLSTM [28]			57.9
He et al. 2023 [55]	PSTNet++ [56]	Point Cloud	TIMo training set (normal data only)	**78.2**

TIMo Anomaly Dataset - Tilted view data				
Method	**Base/Backbone**	**Modalities**	**Training data**	**AUROC**
TIMo Dataset 2022 [20]	CAE* [36]	Depth	TIMo training set (normal data only)	66.4
	ConvLSTM [28]			62.8
Schneider et al. 2022 [37]	R-CAE* [36]	Depth	TIMo training set (normal data only)	70.0
	P-CAE* [36]			**78.1**
	R-ViT-AE‡ [54]			71.7
	P-ViT-AE‡ [54]			71.2
	P-ConvLSTM [28]			67.5
	R-CAE* [36]	IR		65.5
	P-CAE* [36]			64.9
	R-ViT-AE‡ [54]			64.5
	P-ViT-AE‡ [54]			64.2
	P-ConvLSTM [28]			63.8
He et al. 2023 [55]	PSTNet++ [56]	Point Cloud	TIMo training set (normal data only)	**81.0**

and adapting for testing on the real front-seat view TICaM dataset.

Finally, Sharma et al. [51] investigated network architectures such as ShapeConv [52] or Depth-aware CNN [53] and proposed a combination of them with the addition of multi-task learning for depth completion. Most importantly, this work showed that the segmentation performance when RGB+Depth data are used can be matched or even surpassed by an IR+Depth input combination that can be provided by a single ToF camera device (no RGB).

Overall, results on the benchmark show that existing solutions can already achieve very promising performance. Synthetic to real adaptation remains a very relevant problem, considering its viability for a scalable solution towards direct deployment of the network models to commercial vehicle applications.

B. TIMO BENCHMARK

To this day, there are not as many reported results on the TIMo benchmark as there are for TICaM, which is also related to its later time of publication. However, some key aspects mainly concerning unsupervised anomaly detection from ToF images were investigated in depth in [37].

In Table 3, we summarize the results reported on the TIMo Anomaly detection dataset. Results on the top-down and tilted-view data (see Figure 2) are shown separately in two

TABLE 4. Summary of reported experimental results on VIZTA TIMo dataset for the task person detection.

TIMo Person Detection Dataset					
Method	**Base/Backbone**	**Modalities**	**Training data**	**B.Box mAP**	**Mask mAP**
TIMo Dataset 2022 [20]	Mask R-CNN[33] YOLACT [34]	Depth	Full TIMo detection training dataset	92.9 88.6	92.8 93.0

sections of the Table. The evaluation metric used is the area under the ROC curve (AUROC), which is the standard metric for unsupervised anomaly detection.

Initially, as a benchmark baseline for the TIMo anomaly detection, two typical methods for unsupervised anomaly detection were applied to the ToF depth modality in [20]. CAE is a convolutional autoencoder that is trained on normal data to reconstruct the input. ConvLSTM is a convolutional LSTM that is trained to predict the current frame from a sequence of N previous frames. In both cases, a threshold on the reconstruction or prediction accuracy is used to classify frames as anomalous. In [20], both methods showed moderate performance in this challenging task.

In [37], a more detailed evaluation of different aspects of the dataset was performed and the achieved AUROC was significantly increased for the TIMo benchmark. The main reason for this increase was the use of the foreground aware loss as described in Section V-C. In addition to approaches and network architectures based on frame prediction (denoted P-) or frame reconstruction (denoted R-), a vision transformer-based autoencoder network was tested as well (ViT-AE). The highest AUROC scores were however achieved by the CAE networks on both the top-down view and the tilted-view dataset, which could be attributed to the size of the training dataset as well since transformer networks are expected to require more training data.

Interestingly [37] also compared the anomaly detection performance when the IR modality of the ToF camera is used instead of the depth map, and showed the IR performs considerably worse. Other evaluations not included on Table 3, used a splitting of the testing dataset by anomaly type (aggresive behavior, medical emergency, left-behind object) and showed that the aggresive behavior category was the easiest to detect, while the other two categories lacking dynamic motion were more challenging for the tested methods (see results in [37]).

Recently, He et al. [55] presented the top-performing method on the TIMo dataset up to now. In this work, point clouds derived from the depth images were used in multi-frame autoencoder reconstruction method for anomaly detection.

In Table 4 the results on the TIMo person detection benchmark from [20] are summarized. Two networks (Mask R-CNN [33], YOLACT [34]) were evaluated with both showing comparatively good results on the benchmark, indicating that this supervised task is less challenging than the anomaly detection.

VIII. DISCUSSION AND CONCLUSION

In this work, we presented an overview of our work on ToF-based perception within the 42 Months of the VIZTA research project.

The most significant outcome of this work is the introduction of the notion that high-accuracy perception for different tasks is possible with a single ToF camera, profiting from advantages such as privacy preservation, natural foreground separation, indoor lighting variation robustness and facilitation of synthetic training data generation. This is contrast to the highly popular RGB-D combination of a normal camera with a depth sensor that induces additional effort for calibration and synchronization, data processing and overall power consumption. Indeed, in the experimental results section, we refer to examples where the competitiveness of Depth+IR from a single ToF sensor against RGB-D devices was validated.

Equally important outcomes of VIZTA are the shared know-how in the use ToF depth images specifically, as well as the release of two publically avaialble ToF datasets, TICaM and TIMo, that can serve as a reference benchmark in ToF-based algorithm development by the scientific community as well as the industry in a multitude of problems including in-car object and person detection, activity recognition and anomaly detection in smart buildings.

Looking at the benchmark datasets and results reported thus far, it is clear that there is a varying degree of task difficulty. Detection tasks (especially large object / human) have already reached very high accuracy levels on our benchmarks (around 90%) even though some applications such as people counting will require almost perfect accuracy (above 99%). However, in a concrete application like people counting one would complement the person detection algorithm which can lift the performance significant by filtering out transient false detections. Meanwhile, unsupervised anomaly detection is a much more challenging task due to its non-deterministic definition with very promising results shown already on TIMo.

ToF Sensing is expected to increase in popularity, either on its own or in combination with other sensors in the upcoming years. At the same time, it is expected to see more product-level systems that rely on such modalities. The VIZTA project prototype results have already been or are expected to be integrated in commercial systems developed by the industrial project partners.

ACKNOWLEDGMENT

The authors would like to thank all the VIZTA project partners for an excellent collaboration (https://www.vizta-ecsel.eu/) during the project and their DFKI colleagues Hartmut Feld, Jigyasa Singh Katrolia, Pascal Schneider, and Oliver Wasenmüller, who worked as a researcher in the project, also would like to thank the project partner IEE S. A.—in particular to Frédéric Grandidier for their close collaboration in data definition and recording and Steve Dias da Cruz for the collaboration in synthetic data generation, also would like to thank all the students and colleagues that supported them in the recording and annotation of data (Raisul Islam, Sai Srinivas Jeevanandam, Ahmed Elsherif, Iuliia Brishtel, Phyllis Qian, Lars Kraemer, Ibrahim Abdelaziz, Dennis Stumpf, Brijesh Varsani, and Kannan Balakrishnan), and also would like to thank all the colleagues and friends who acted as subjects in their datasets.

REFERENCES

[1] V. Blahnik and O. Schindelbeck, "Smartphone imaging technology and its applications," *Adv. Opt. Technol.*, vol. 10, no. 3, pp. 145–232, Jun. 2021, doi: 10.1515/aot-2021-0023.

[2] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes," in *Proc. 11th Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2013, pp. 548–562.

[3] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon, "KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera," in *Proc. 24th Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2011, pp. 559–568.

[4] *Microsoft Azure Kinect*. Accessed: Mar. 29, 2023. [Online]. Available: https://azure.microsoft.com/en-us/products/Kinect-dk/

[5] *Intel Realsense*. Accessed: Jul. 25, 2023. [Online]. Available: https://www.intelrealsense.com/

[6] *VIZTA Project EU ECSEL*. Accessed: Jul. 25, 2023. [Online]. Available: https://www.vizta-ecsel.eu/

[7] M. Tölgyessy, M. Dekan, L. Chovanec, and P. Hubinský, "Evaluation of the Azure Kinect and its comparison to Kinect v1 and Kinect v2," *Sensors*, vol. 21, no. 2, p. 413, Jan. 2021.

[8] R. Lange, P. Seitz, A. Biber, and S. Lauxtermann, "Demodulation pixels in CCD and CMOS technologies for time-of-flight ranging," *Proc. SPIE*, vol. 3965, pp. 177–188, May 2000.

[9] C. Tubert et al., "4.6μm low power indirect time-of-flight pixel achieving 88.5% demodulation contrast at 200 MHz for 0.54 MPix depth camera," in *Proc. IEEE 47th Eur. Solid State Circuits Conf. (ESSCIRC)*, Sep. 2021, pp. 135–138, doi: 10.1109/ESSCIRC53450.2021.9567878.

[10] O. Wasenmüller and D. Stricker, "Comparison of Kinect V1 and V2 depth images in terms of accuracy and precision," in *Proc. Int. Workshops Comput. Vis.*, in Lecture Notes in Computer Science, vol. 10117, C. Chen, J. Lu, and K. Ma, Eds. Cham, Switzerland: Springer, 2016, pp. 34–45, doi: 10.1007/978-3-319-54427-4_3.

[11] *Smart Buildings*. Accessed: Jun. 9, 2023. [Online]. Available: https://iee-sensing.com/en/building-management-security.html

[12] (2017). *SoftKinetic's Gesture Control Technology Rolls Out in Additional Car Model*. Accessed: Aug. 4, 2023. [Online]. Available: https://www.prnewswire.com/news-releases/softkinetics-gesture-control-technology-rolls-out-in-additional-car-model-300461231.html/

[13] (2020). *The Technology of the New Mercedes-Maybach S-Class*. Accessed: Aug. 4, 2023. [Online]. Available: https://media.mercedes-benz.com/article/880466b3-bed4-41fa-bda2-6041b45c53c3/

[14] *Analog Devices 3D Time of Flight (ToF)*. Accessed: Jun. 9, 2023. [Online]. Available: https://www.analog.com/en/applications/technology/3d-time-of-flight.html

[15] *Time of Flight Image Sensor*. Accessed: Jun. 9, 2023. [Online]. Available: https://www.sony-semicon.com/en/products/is/industry/tof.html

[16] H. Feld, B. Mirbach, J. Katrolia, M. Selim, O. Wasenmüller, and D. Stricker, "DFKI cabin simulator: A test platform for visual in-cabin monitoring functions," in *Commercial Vehicle Technology 2020/2021*. Cham, Switzerland: Springer, 2021, pp. 417–430.

[17] *Optitrack Motion Capture*. Accessed: Mar. 29, 2023. [Online]. Available: https://optitrack.com/

[18] *Opends Driving Simulator*. Accessed: Mar. 29, 2023. [Online]. Available: https://opends.dfki.de/

[19] J. S. Katrolia, A. Elsherif, H. Feld, B. Mirbach, J. Rambach, and D. Stricker, "TICaM: A time-of-flight in-car cabin monitoring dataset," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2021, pp. 1–7.

[20] P. Schneider, Y. Anisimov, R. Islam, B. Mirbach, J. Rambach, D. Stricker, and F. Grandidier, "TIMo—A dataset for indoor building monitoring with a time-of-flight camera," *Sensors*, vol. 22, no. 11, p. 3992, May 2022.

[21] D. Stumpf, S. Krauß, G. Reis, O. Wasenmüller, and D. Stricker, "SALT: A semi-automatic labeling tool for RGB-D video sequences," 2021, *arXiv:2102.10820*.

[22] M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. ReiB, M. Voit, and R. Stiefelhagen, "Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2801–2810.

[23] Y. Abouelnaga, H. M. Eraqi, and M. N. Moustafa, "Real-time distracted driver posture classification," 2017, *arXiv:1706.09498*.

[24] A. Jain, H. S. Koppula, S. Soh, B. Raghavan, A. Singh, and A. Saxena, "Brain4Cars: Car that knows before you do via sensory-fusion deep learning architecture," 2016, *arXiv:1601.00740*.

[25] E. Ohn-Bar, S. Martin, A. Tawari, and M. M. Trivedi, "Head, eye, and hand patterns for driver activity recognition," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 660–665.

[26] S. D. Da Cruz, O. Wasenmüller, H.-P. Beise, T. Stifter, and D. Stricker, "SVIRO: Synthetic vehicle interior rear seat occupancy dataset and benchmark," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 962–971.

[27] M. Selim, A. Firintepe, A. Pagani, and D. Stricker, "AutoPOSE: Large-scale automotive driver head pose and gaze dataset with deep head orientation baseline," in *Proc. 15th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2020, pp. 599–606.

[28] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545.

[29] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6479–6488.

[30] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[31] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 168–172.

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, vol. 28, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

[33] 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.

[34] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9156–9165.

[35] J. Katrolia, L. Krämer, J. Rambach, B. Mirbach, and D. Stricker, "An adversarial training based framework for depth domain adaptation," in *Proc. 16th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2021, pp. 353–361.

[36] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733–742.

[37] P. Schneider, J. Rambach, B. Mirbach, and D. Stricker, "Unsupervised anomaly detection from time-of-flight depth images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 230–239.

[38] M. Braham, A. Lejeune, and M. Van Droogenbroeck, "A physically motivated pixel-based model for background subtraction in 3D images," in *Proc. Int. Conf. 3D Imag. (IC3D)*, Dec. 2014, pp. 1–8.

[39] J. S. Katrolia, L. Krämer, J. Rambach, B. Mirbach, and D. Stricker, "Semantic segmentation in depth data: A comparative evaluation of image and point cloud based methods," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 649–653.

[40] *NVIDIA Jetson AGX Xavier*. Accessed: Jul. 26, 2023. [Online]. Available: https://www.nvidia.com/de-de/autonomous-machines/embedded-systems/jetson-agx-xavier/

[41] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[42] J. Long, E. Shelhamer, and T. Darrell, ''Fully convolutional networks for semantic segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[43] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, ''PointNet: Deep learning on point sets for 3D classification and segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 77–85.

[44] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, ''PointNet++: Deep hierarchical feature learning on point sets in a metric space,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5105–5114.

[45] N. Ebert, P. Mangat, and O. Wasenmuller, ''Multitask network for joint object detection, semantic segmentation and human pose estimation in vehicle occupancy monitoring,'' in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2022, pp. 637–643.

[46] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, ''CSPNet: A new backbone that can enhance learning capability of CNN,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1571–1580.

[47] S. Dias Da Cruz, B. Taetz, T. Stifter, and D. Stricker, ''Autoencoder and partially impossible reconstruction losses,'' *Sensors*, vol. 22, no. 13, p. 4862, Jun. 2022.

[48] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*.

[49] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[50] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, ''Densely connected convolutional networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[51] P. Sharma, J. S. Katrolia, J. Rambach, B. Mirbach, D. Stricker, and J. Seiler, ''Achieving RGB-D level segmentation performance from a single ToF camera,'' 2023, *arXiv:2306.17636*.

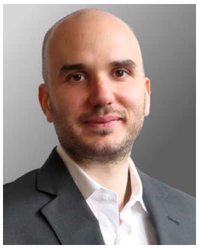
[52] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li, ''ShapeConv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7068–7077.

[53] W. Wang and U. Neumann, ''Depth-aware CNN for RGB-D segmentation,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 135–150.

[54] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and J. Uszkoreit, ''An image is worth 16×16 words: Transformers for image recognition at scale,'' in *Proc. Int. Conf. Learn. Represent.*, 2021.

[55] T. He and W. Wang, ''Point cloud video anomaly detection based on point spatio-temporal auto-encoder,'' 2023, *arXiv:2306.04466*.

[56] H. Fan, X. Yu, Y. Yang, and M. Kankanhalli, ''Deep hierarchical representation of point cloud videos via spatio-temporal decomposition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9918–9930, Dec. 2022.



JASON RAMBACH received the Ph.D. degree in computer science from the University of Kaiserslautern, in 2020.

His Ph.D. dissertation titled ''Learning Priors for Augmented Reality Tracking and Scene Understanding.'' He has been with DFKI Augmented Vision, Kaiserslautern, since 2015. Currently, he is a Senior Researcher leading the team ''Spatial Sensing and Machine Perception'' of approximately ten researchers working on depth sensing and geometric/semantic scene understanding using machine learning. His research interests include SLAM and semantic scene understanding, object pose estimation and tracking, anomaly detection, hybrid AI, robotic vision, and augmented reality. He has over 40 publications in leading computer vision and augmented reality conferences, the Best Paper Award from the IEEE International Symposium on Mixed and Augmented Reality (ISMAR), in 2017, and two awards at the BOP Object Pose Estimation Challenge, in 2022, at ECCV. He is a Reviewer of several scientific journals and conferences (CVPR, ECCV, ICRA, WACV, and BMVC). Since 2022, he has been a Coordinator of the EU Horizon Project HumanTech, applying AI in the construction industry for Scan2BIM, wearables, and assistance robots.

BRUNO MIRBACH received the Dr.rer.nat. degree in theoretical physics from the University of Kaiserslautern, in 1996.

He was a Postdoctoral Researcher with the Center for Nonlinear and Complex Systems, Como, Italy, the Max-Planck-Institute of the Physics of Complex Systems, Dresden, Germany, and the University of Ulm, Germany, with a research focuses on non-linear dynamics and quantum chaos. In 1999, he joined automotive industry working on the applied research and development of intelligent optical sensor systems for environment perception and occupant monitoring. At two automotive suppliers, he has been leading teams dedicated to the research and development of computer vision and machine learning algorithm, with numerous publications and patents in 3D-vision, 2D/3D sensor fusion, and machine learning. Since 2019, he has been a part-time Senior Researcher with the German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany. In the Augmented Vision Department, he is mainly contributing to European research projects (Horizon Europe Program) and industry projects, with a main research focuses on 3D-vision and deep learning.

YURIY ANISIMOV received the M.Eng. degree in electrical engineering from the Moscow State University of Mechanical Engineering (MAMI), with the thesis topic of ''Road lane and road signs recognition system'' and the Ph.D. degree from DFKI, in 2015. During his studies, he completed an internship with South Korean Hyundai motor plant and the University of Ulsan for the automotive engineering and robotics topics. He joined DFKI as a Researcher, where he was involved in several research projects. His work focused on real-time three-dimensional reconstruction on embedded hardware and deep-learning based object detection from ToF depth images on edge devices and deep-learning-driven photometric stereo. His research interests include parallel programming, real-time algorithms, embedded hardware, three-dimensional reconstruction and its applications, teaching, co-organizing the exercises on ''2D image processing'' course, projects and seminars on ''3D computer vision'' and ''2D image processing'' topics, and being a Lecturer in ''advanced topics in computer vision and deep learning'' course with Rheinland-Pfälzische Technische Universität.

DIDIER STRICKER is currently a Professor in computer science with Rheinland-Palatinate Technical University (RPTU) and the Scientific Director of the German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern, where he leads the research department ''Augmented Vision.'' Between 2011 and 2016, he acted as the Honorary CEO of International Network, GraphicsVision.AI. His research interests include cognitive interfaces, user monitoring and on-body-sensor-networks, computer vision, video/image analytics, and human–computer interaction.

He received the Innovation Prize of the German Society of Computer Science, in 2006. He organized as the General Chair the first ''2002 IEEE&ACM International Symposium on Mixed and Augmented Reality'' (ISMAR), in Darmstadt, Germany, and was a member of the Steering Committee, from 2000 to 2007. He got several award for best papers or demonstrations at different conferences. He registered several patents on tracking and augmented reality. He serves as a reviewer for different European or national research organizations. He is a reviewer of different journals and conferences in the areas of VR/AR and computer vision.