Accurate detection and 3D localization of humans using a novel YOLO-based RGB-D fusion approach and synthetic training data

Timm Linder[1,2] Kilian Y. Pfeiffer[3*] Narunas Vaskevicius[1*] Robert Schirmer[1] Kai O. Arras[1]

***Abstract*— While 2D object detection has made significant progress, robustly localizing objects in 3D space under presence of occlusion is still an unresolved issue. Our focus in this work is on real-time detection of human 3D centroids in RGB-D data. We propose an image-based detection approach which extends the YOLO v3 architecture with a 3D centroid loss and mid-level feature fusion to exploit complementary information from both modalities. We employ a transfer learning scheme which can benefit from existing large-scale 2D object detection datasets, while at the same time learning end-to-end 3D localization from our highly randomized, diverse synthetic RGB-D dataset with precise 3D groundtruth. We further propose a geometrically more accurate depth-aware crop augmentation for training on RGB-D data, which helps to improve 3D localization accuracy. In experiments on our challenging intralogistics dataset, we achieve state-of-the-art performance even when learning 3D localization just from synthetic data.**

I. INTRODUCTION

Detection of persons and objects in 3D space is an important capability for service, domestic and industrial robots that interact with their environment. In indoor scenarios, RGB-D sensors such as the Kinect v2 are often used for this purpose. However, while recent advances in computer vision have mostly solved the unimodal 2D detection problem on RGB images, it is not yet fully understood what is the best representation and strategy for approaching the 3D detection task, especially on multimodal RGB-D data, where large-scale datasets are scarce and we want to benefit as much as possible from existing work on 2D detection.

In this paper, we tackle the problem of learning to detect and *accurately localize 3D centroids* in RGB-D data in an end-to-end fashion, with an experimental focus on human detection in a challenging intralogistics context. We show that 3D localization can, to a large part, be learned from a diverse and highly randomized synthetic RGB-D dataset with perfect 3D groundtruth, and that for successful fine-tuning on real-world data, no manual 3D annotation is required. Our proposed real-time approach uses a strong image-based YOLO v3 single-stage detector as starting point, extends the RGB feature extractor with a separate depth stream via mid-level fusion, and utilizes a hardwired transfer learning strategy that can reuse existing pretrained weights from large-scale 2D object detection datasets. Thereby, incorporating the depth channel does not require training from scratch and thus does not lead to a loss in 2D detection performance. For 3D localization, we extend the resulting RGB-D YOLO v3 detector with a centroid regression output. Finally, we propose a depth-aware, scale-preserving variant of zoom-in/zoom-out training-time augmentation [1] for accurate depth regression.

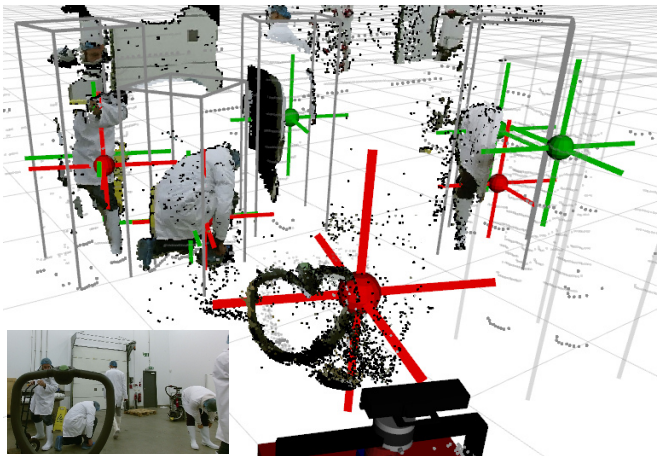


Fig. 1: Our method (green) localizes 3D person centroids much more robustly than a baseline (red) on our intralogistics dataset.

As opposed to the existing methods we compare against [2]–[5], our end-to-end 3D regression can exploit complementary RGB and depth information by fusing modalities at the feature extractor stage, and does not rely on any 3D point cloud representation. It is therefore robust to missing depth data and works well under partial occlusion.

Our key contributions in this paper are:

1) We demonstrate that accurate 3D localization under partial occlusion is an unsolved issue, which is an important aspect *e.g.* for human detection in robotics.
2) We are, to our best knowledge, the first to propose an RGB-D fusion strategy for the fast YOLO v3 one-stage detector, with an accompanying transfer learning strategy that leverages existing large-scale 2D datasets.
3) Via heavy domain randomization, we are able to learn end-to-end regression of 3D human centroids from a synthetically rendered multi-person RGB-D dataset.
4) We find that standard 2D crop/expansion augmentations are unsuitable for depth data, and propose a geometrically more accurate variant that accounts for the resulting shift of focal length.
5) On our challenging real-world RGB-D dataset from the intralogistics domain, our method outperforms existing baselines in 3D person detection without requiring additional hand-annotated 3D groundtruth for training.

Our approach achieves real-time speed at 25 Hz on a GPU.

[1]Robert Bosch GmbH, Corporate Research, Stuttgart, Germany.
✉ timm.linder@de.bosch.com
[2]Department of Computer Science, University of Freiburg, Germany.
[3]RWTH Aachen University. Research conducted as an intern at Bosch.
* The second and third author contributed equally to this work.

Method	Modalities	Detector	Fusion strategy	Output	Dataset
Munaro *et al*., 2014 [6]	RGB+D	PCL+HOG-SVM	3D proposals → 2D classifier	3D boxes	KTP
Mees *et al*., 2016 [7]	RGB+D+Flow	Fast R-CNN	2D late (adaptive gating)	2D boxes	InOutDoor (IO)
Vasquez *et al*., 2017 [2]	RGB+D	Fast R-CNN	3D proposals → 2D detector	2.5D centroids	MobilityAids, IO
Guerry *et al*., 2017 [8]	RGB+D	Faster R-CNN	2D early/mid/late	2D boxes	Onera, Mensa, IO
Ali *et al*., 2018 [9]	Projected LiDAR	YOLO v2	–	3D boxes + orient.	KITTI
Simon *et al*., 2018 [10]	Projected LiDAR	YOLO v2	–	3D boxes + orient.	KITTI
Qi *et al*., 2018 [11]	RGB+D/LiDAR	FPN+Fast R-CNN	2D boxes → 3D frustums	3D boxes + orient.	KITTI, SUN
Zimmermann *et al*., 2018 [4]	RGB+D	OpenPose	2D joints → 3D voxel grid	3D body joints	MKV-t, CAP-t
Lewandowski *et al*., 2019 [12]	D	FPFH-SVM	–	3D boxes	Supermarket
Kollmitz *et al*., 2019 [3]	RGB or D	Faster R-CNN	–	3D centroids	MobilityAids
Ophoff *et al*., 2019 [13]	RGB+D	YOLO v2	2D mid-level features	2D boxes	KITTI, EPFL
Our approach	RGB+D	YOLO v3	2D mid-level features	3D centroids	Intralogistics

TABLE I: Qualitative comparison of related work on person detection in 3D and RGB-D

II. RELATED WORK

A. 3D person detection in RGB-D

There is a vast amount of literature on multi-modal [14] and RGB-D-based [15] object recognition. In Table I we list recent works that were evaluated on human detection. Some fuse color and depth information, but only output 2D bounding boxes and do not tackle the issue of 3D localization [7], [8], [13]. Several approaches utilize a geometric 3D point cloud representation [2], [4], [6], [11], [12], which comes with certain drawbacks. For example, the method by Qi *et al*. [11] suffers from three weaknesses towards which our proposed method should be more robust: 1.) Their 3D stage fails to accurately localize objects in locally sparse point clouds. Here, our approach can leverage complementary RGB data as it does not rely on a point cloud representation. 2.) When multiple instances of a class share the same 3D frustum, only a single instance is detected. This scenario is frequent in our indoor environments, where humans often partially occlude each other. 3.) Their RGB-based 2D detector fails under difficult lighting conditions, where our method can exploit complementary depth data due to our mid-level fusion strategy. The methods by Munaro *et al*. [6] and Vasquez *et al*. [2], which we include as baselines in our experiments, suffer from similar conceptual limitations.

More recent works therefore often leverage a 2D image-based representation [3], [7]–[10], [13] in order to exploit advances in 2D object detection. They are based upon single-stage detectors like YOLO v2 [16] or the computationally more expensive two-stage R-CNN framework [17], [18].

Most closely related to our work are the methods by Ophoff *et al*. [13] and Kollmitz *et al*. [3]. With focus only on 2D detection, [13] incorporates RGB+D fusion into the earlier YOLO v2 architecture. It is not as deep, uses no shortcut connections and does not include a feature pyramid compared to the YOLO v3 [19] architecture that our work is based upon, which imposes extra constraints on where we can fuse features. Similar to our method, and contrary to their earlier work [2], Kollmitz *et al*. [3] do not employ a 3D point cloud representation, and instead utilize an end-to-end 2D detector with 3D centroid output. Their two-stage approach is evaluated in a multi-class hospital scenario including persons with walking aids. They provide separate models for detection on either RGB or depth data. In contrast, our method follows an efficient one-stage approach, performs principled fusion of the RGB+D modalities to exploit complementary information, utilizes synthetic training data to learn 3D localization, and incorporates a depth-aware crop augmentation scheme that improves 3D localization. We evaluate our method and baselines on a novel, challenging intralogistics dataset.

B. Learning from synthetic RGB-D data

Learning from simulation and transfer into the real world are currently quickly evolving topics in computer vision and robotics. Most work so far focuses on rigid objects. [20] explore domain randomization for robotic manipulation. Using synthesized RGB images with random camera and object positions, lighting, and textures, they learn accurate 3D detectors for geometric primitives without pretraining on real images. [21] learn to estimate 3D orientation of objects using synthetic RGB images based upon CAD models with domain randomization. [22] focus on category-agnostic 2D instance segmentation using synthetic depth also from CAD.

Humans vary greatly in shape and appearance, and are thus particularly challenging to simulate. The work by Shotton *et al*. [23] on articulated human pose estimation focused on single-person scenarios using synthetic depth images, without simulating any 3D background. This also applies to the SURREAL dataset [24], which however contains additional modalities such as RGB. [25] perform semantic segmentation on KITTI [26] by training on synthetic RGB images and groundtruth masks from a commercial computer game. In contrast to these works, our synthetic dataset [27] focuses on multi-person human detection in clutter and under occlusion. It contains up to several dozens of person instances per frame, diverse 3D backgrounds and large amounts of foreground occluder objects with strong domain randomization (unlike [28]–[30]). It is composed of synchronized RGB-D pairs, where we additionally model noise characteristics of the Kinect v2 time-of-flight sensor.

III. METHOD

We now present our solution for robust detection and 3D localization of persons from RGB-D data. Following the data flow, we first describe our synthetic RGB-D dataset that we use to learn 3D human detection and localization. We then propose a depth-aware and scale-preserving augmentation scheme for training 3D detectors on RGB-D data. Finally, we present our modifications to the YOLO v3 detector [19]

Fig. 2: 3D groundtruth joint locations on our synthetic RGB-D and our real-world RGB-D datasets. The latter ones are derived from offline 3D human pose estimation [4], and only used for fine-tuning on real-world data if desired. Our video shows further examples.

to fuse RGB and depth information, and regress 3D centroids in an end-to-end fashion along with a training scheme that allows us to benefit from existing 2D detection datasets.

A. Synthetic RGB-D data for learning 3D localization

Obtaining a sufficiently diverse RGB-D dataset with accurate 3D groundtruth is difficult and time-consuming in the real world. We therefore propose to learn 3D human localization from a synthetic RGB-D dataset that has been rendered using a semi-photorealistic game engine (Unreal Engine 4). Our initial work in this direction [27] focused on 2D detection. We now extended our simulation to output sensor-centric groundtruth 3D coordinates and visibility flags for 23 human body joints (Fig. 2, left). We increased the number of rigged 3D human models by a factor of 4 and doubled the amount of motion-capture animations to around 100 each, now also including sitting, kneeing and lying poses. Inspired by the success of 2D synthetic occlusion augmentation [31], we significantly increased the number of 3D occluder objects to over 700 to enhance foreground diversity (some of which can be seen in Fig. 3, left).

For each of the 6 scenes from [27], we initially generate 5,000 RGB-D frames. In [27], we showed that appropriate filtering of groundtruth bounding boxes is important for successful training: We therefore set ignore flags on all 2D groundtruth person boxes with extremely low contrast, with a groundtruth instance mask that covers less than 300 square pixels, or where more than 80% of the essential body joints are either truncated or occluded after projecting them onto the 2D instance mask. We then iterate over the 5,000 frames per scene, remove all frames which have less than two remaining (non-ignore) boxes, followed by random subsampling to finally obtain 2,500 sufficiently dense frames per scene. Combining all six scenes, the resulting RGB-D dataset thus consists of 15,000 training samples.

B. Weak real-world labels from 3D human pose estimation

[3] propose a relatively robust clustering-based heuristic to derive groundtruth 3D centroid coordinates for their training set, without requiring manual 3D annotation. This approach can fail if persons are truncated, for example when only the head or an arm are visible. Unless such persons are skipped, which can prevent difficult examples from ending up in the training set, the centroid would be offset to the top or to the side. We therefore propose 1.) to use a more informed approach, by leveraging offline 3D human pose estimation [4] to derive weak groundtruth from predicted 3D body joints, 2.) to select a fixed, central body joint as the 'centroid' regression target, which is more stably attached to the human body under truncation or unsual poses (*e.g.* stretching out a single arm) and thus more suitable for 3D human tracking [32] or 3D articulated pose estimation applications. For the latter, top-down methods such as [33] often require the pelvis joint, localized between the hip joints, as body-centric root joint for input, which we adapt as 3D regression target. However, in principle, our method can be trained on any (derived) body joint, as shown in Fig. 2 (right).

C. Depth-aware augmentation

The 2D data augmentation pipeline of our underlying YOLO v3 implementation [34] involves random cropping and expansion of the image with corresponding adjustment of the bounding boxes, originally referred to as "zoom in" and "zoom out" [1]. This is followed by resizing to a fixed-size square input image provided to our network during training.

Directly applying crop or expansion augmentation and resizing to an RGB-D image can distort the understanding of objects' metric scale in the perceived environment, which is essential for accurate 3D perception. Therefore, we propose a depth-aware variant of this augmentation that adapts groundtruth depth labels and input depth to the current "zoom level" at training time, and preserves the metric scale and aspect ratio of the original image for a physically more well-grounded representation.

Our network has a square input resolution of $d_n \times d_n$ pixels during training. Therefore, to preserve aspect ratio, we constrain ourselves to random square crops of varying size $d_c \times d_c$. Under the assumption of a single sensor with known camera matrix $\mathbf{K}$, resizing of a crop to the input dimension $d_n \times d_n$ can be expressed as a zooming operation [35] with zoom factor $s = d_n/d_c$. Zooming is usually attributed to a change in focal length. Instead, to keep the intrinsic parameters intact, we apply the scaling to the depth values:

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \mathbf{K} \begin{pmatrix} s & & \\ & s & \\ & & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \frac{1}{z} = \mathbf{K} \begin{pmatrix} x \\ y \\ \frac{z}{s} \end{pmatrix} \frac{1}{\frac{z}{s}} \qquad (1)$$

here (x, y, z) is a 3D point resolved in the RGB-D sensor frame and z/s is the new scaled depth at input pixel (u, v). While this operation is not physically well-grounded for arbitrary crops not centered at the principal point of the RGB-D sensor, this approximation already yields a significant improvement, as shown later in our ablation studies. In addition to depth scaling used during training, we scale input depth measurements at inference time according to the resize transformation applied to the RGB-D image.

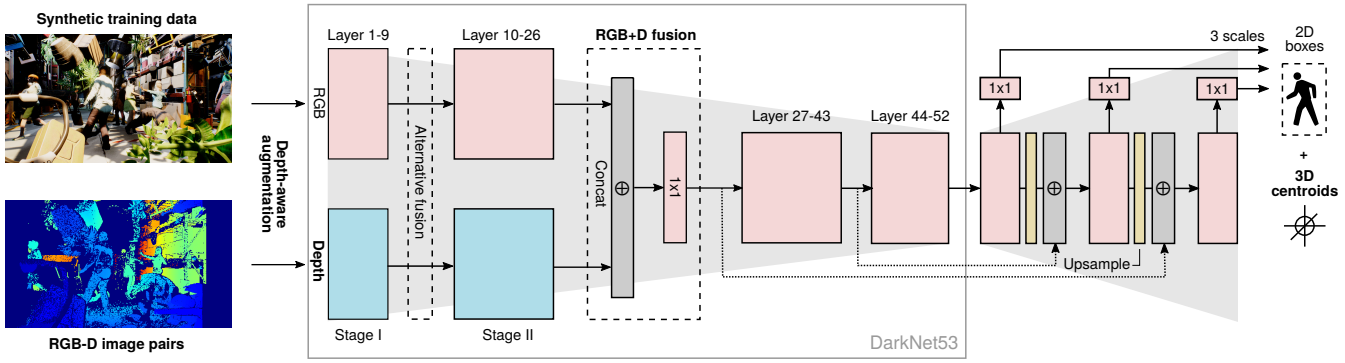


Fig. 3: Overview of our proposed approach, which extends the YOLO v3 [19] detector with mid-level RGB+D feature fusion, depth-aware augmentation, and 3D centroid regression. We show that the latter can be learned from synthetic RGB-D images.

D. Network architecture

Our method is based upon the YOLO v3 network, which we modified to also predict 3D centroids. First, to leverage depth information, we extend the Darknet-53 backbone to accommodate the additional single-channel depth data. Therefore, we duplicate layers up until a fusion point resulting in an RGB and depth specific backbone (blue in Fig. 3).

We evaluate two different mid-level fusion points, which are placed at the end of residual stages in the Darknet-53 architecture, in our case after Layer 9 or Layer 26. Mid-level fusion has shown to lead to good results [13], [36]. Furthermore, we fuse the modalities before the pyramid structure of the network begins (see Fig. 3). The modalities are fused by concatenating the output features of both backbones along the channel dimension and using a 1×1 convolution to halve the number of channels to the original channel dimension.

The final 1×1 convolutions of the three output stages are extended to predict a centroid (c_x, c_y, c_z) for each anchor box. c_z is regressed directly in metric scale. The c_x and c_y coordinate are first predicted in image coordinates (c_u, c_v) and then backprojected with help of c_z and the camera matrix. This mirrors the unit system of our input. Furthermore, we formulate the regression targets t_{c_u}, t_{c_v} in a constraint manner, relative to the bounding box coordinates where (b_u, b_v) is the top-left corner of the bounding box with height b_h and width b_w in pixel coordinates:

$$\begin{aligned} c_u &= b_u + b_w \sigma(t_{c_u}) \\ c_v &= b_v + b_h \sigma(t_{c_v}) \\ c_z &= t_{c_z} \end{aligned} \tag{2}$$

This limits the centroid to lay within the predicted bounding box. The 2D YOLO loss per anchor box is then extended by an additional term

$$\mathcal{L}_{\text{centroid}} = |t_{c_z} - \hat{t}_{c_z}| + \sum_{i \in \{u,v\}} \text{BCE}(\sigma(t_{c_i}), \sigma(\hat{t}_{c_i})) \tag{3}$$

where $\hat{t}_{c_i}$ denotes the groundtruth label. While for c_u, c_v we keep the sigmoid binary cross-entropy loss used for 2D bounding box centers in our YOLO v3 implementation, we found that ℓ_1 loss works best for centroid depth c_z.

E. Transfer-learning strategy

Our transfer learning strategy is inspired by Ophoff *et al.* [13], but without an extra step to first train a depth-only detector. To benefit from existing large-scale 2D object detection datasets, we first initialize all layers from existing YOLO v3 RGB detector weights pretrained on ImageNet and MS COCO [37], [38]. While other transfer learning strategies, such as proposed in [13], [39], could also be used, we already obtained good results using this approach. For the depth backbone, we duplicate RGB weights, but initialize the first layer (that takes single-channel depth images) from scratch. As indicated by the coloring in Figure 3, the fusion block is initialized using a hardwired fusion scheme such that at the start of training, the existing RGB features are forwarded as-is. In the 3 output layers that we extended with 3D centroid regression, we randomly initialize weights for the new outputs, while leaving the original 2D detection weights unchanged. This initialization strategy, which is illustrated further in our video, allows to maintain the pretrained 2D performance despite the changes to the network architecture.

IV. EXPERIMENTAL SETUP

Our implementation is based upon MxNet using YOLO v3 from GluonCV [34]. We train for a total of 80 epochs using stochastic gradient descent. During a warmup phase of 20 epochs, we gradually increase the learning rate to 6e-4, after which the learning rate is reduced to 1e-6 over 50 epochs using cosine decay [40], [41]. Finally, the model is trained for another 10 epochs at a constant rate of 1e-6. For training, we use Volta V100 GPUs, and a Titan RTX for testing.

A. Real-world intralogistics RGB-D dataset

Our application use-case is person detection in the intralogistics domain, with the goal of making autonomous guided vehicles (AGVs) human-aware. Human detection in such professional environments brings up certain challenges, such as people wearing special clothing; human forklift operators standing on the footrests of their vehicles; narrow, cluttered spaces with significant occlusion, which the robot observes from an ego-centric perspective; and the lack of publicly available datasets, especially for sensor modalities containing sensitive information, such as RGB-D. To train and evaluate our method, we therefore recorded a diverse dataset using

two different AGV platforms equipped with a Kinect v2 sensor at around 1.50m and 1.80m height. Data has been recorded over several weeks at four different locations (two warehouses, a small food factory, and a robotics laboratory with forklifts and warehouse shelves). It includes scenes with very few people, as well as very crowded scenes with up to around 20 people that frequently occlude each other and have very similar appearance due to wearing protective clothing.

From these recordings, we selected around 3.1k diverse frames and split them into a training set of 1.5k, a validation set of 0.5k and a 2D test set of 1.1k real-world frames, with each split recorded at a different location or day. Each frame consists of a registered pair of RGB-D images, where we manually annotated 2D person bounding boxes. On the training set, we derive weak 3D groundtruth as described in Section III-B. For 3D evaluation, we labeled a continuous 60-second test set sequence from one of the environments with 3D centroids, using our trajectory-based annotation tool [32] which we extended for annotation of centroid heights over ground. In the sequence, persons are highly dynamic, assume different poses, and some push carts around. For evaluation, we mark all centroids with ignore flags that are too heavily occluded in the point cloud, or outside of the Kinect v2 depth camera's field of view (with 8m range limit), in order not to penalize baseline methods which rely on the availability of depth data.

B. Baselines

We compare our approach with 5 different RGB-D baselines. Besides [2]–[4], [6] (see Table I), we also include an RGB-only YOLO v3 baseline that naïvely lifts 2D centroids into 3D by computing median depth [5]. The 2D detector for this method was trained on MS COCO; with naïve lifting into 3D, we saw no improvement in 3D performance when fine-tuning on our dataset. For [6], the HOG-SVM was trained on a person dataset recorded in an airport environment [32]. We fine-tuned [3] on our real-world training set, as the RGB variant initially did not perform well in our scenario. For both [2] and [3], we show the better variant of RGB or depth, and obtained best results when considering only detections for the person class; we do not use tracking. For the other methods, we use the original, trained model from the publicly available implementations. [4] is not fast enough for our use-case (1-2 Hz), but we still decided to include it as we were interested in how a radically different, bottom-up 3D pose estimation approach would perform on the 3D person detection task. For evaluation, we derive 3D centroids from hip joint positions as described in Sec. III-B; if hips are not detected, we fall back to the median of essential body joints (excluding eyes).

C. Evaluation metrics

For 3D evaluation on our real-world test sequence, we use a modified variant of COCO metrics [38], where instead of bounding box IoU we apply a metric distance threshold and compute 3D average precision (AP) as well as peak-F1 score. For methods that do not output robust estimates of centroid height [2], [3], we are more lenient and perform detection-to-groundtruth association only on ground plane coordinates while ignoring the height. For ablation studies on our precise, synthetic 3D groundtruth, we also report root mean square error (RMSE), as well as 2D bounding box AP according to PASCAL VOC criteria [42] at 0.5 IoU.

Variant	Modality	2D AP VOC	3D AP ↑ 0.25m	3D AP ↑ 0.5m	RMSE ↓
COCO model	RGB	71.0	–	–	–
After finetuning	RGB	74.9	–	–	–
+Centroid regression	RGB	72.5	17.5	48.2	56.9
+Depth-aware augm.	RGB	75.4	46.5	73.4	39.3
After finetuning	RGB+D	75.4	–	–	–
+Centroid regression	RGB+D	73.4	47.6	74.5	34.7
+Depth-aware augm.	RGB+D	76.6	60.6	81.5	30.8
/ Fusion after stage I	RGB+D	76.5	58.9	80.4	31.8
+More synth. data	RGB+D	77.0	66.4	83.0	27.8

TABLE II: Ablation studies on our synthetic validation set with perfect 3D groundtruth. RGB+D fusion after stage II unless noted.

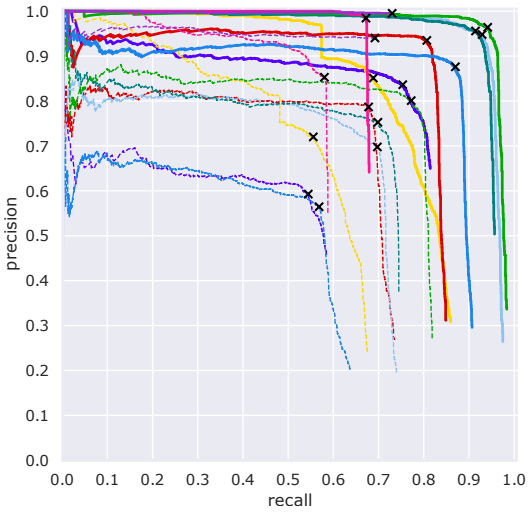
V. RESULTS

A. Quantitative evaluation

Table II shows results of ablation studies on our synthetic validation set (2 extra scenes, 5k diverse frames) with precise groundtruth. We used half of the synthetic training set (7.5k frames) for training. It can be seen that our depth-aware augmentation scheme boosts accuracy in both 2D and 3D significantly, compared to when using standard rectangular crop augmentation and resizing without scaling groundtruth labels and input depth appropriately. With the full synthetic training set, especially 3D localization at the smaller distance threshold improves. Incorporating RGB+D fusion leads to significantly higher 3D accuracy, and improves 2D slightly.

In Table III and the corresponding precision/recall curves, we compare variants of our method to other baselines on a 60-second hand-annotated sequence (1.8k frames) from our real-world intralogistics test set. It can be seen that our RGB-D method fine-tuned only on our real-world data (■) does not achieve a better peak-F1 score than our naïve YOLO v3 baseline trained on MS COCO (■) under the 3D metric with $d = 0.5\text{m}$, and performs especially bad at d=0.25m. This could indicate that our real-world training set is too small. Training only on synthetic images (■), where through domain randomization we can generate an unlimited amount of frames and here thus have 10x more data with precise 3D groundtruth, improves performance drastically at both thresholds. Adding real to the synthetic training data (■) further improves performance. The corresponding RGB-only model (■) is similarly strong at $d = 0.5\text{m}$. However, at d=0.25m, the combination of both modalities is around +9% better in peak-F1, which shows that our network can exploit depth information for more accurate 3D localization. Yet, as we also observe qualitatively, our approach degrades gracefully when only RGB data is available.

By combining 3D clustering-based region proposals with a modern deep learning-based 2D detector, Vasquez *et al.* (■) achieve very good localization accuracy and outperform our method in peak-F1 at $d = 0.25\text{m}$. Here, our method might be at slight disadvantage because we train on pelvis



Method	3D AP ↑		Peak-F1 ↑	
	0.25m	0.5m	0.25m	0.5m
Munaro[et al] [6] (RGB+D)	56.0	77.9	62.7	76.1
Vasquez[et al] [2] (RGB+D, RGB)	66.3	73.1	**79.7**	84.2
Kollmitz[et al] [3] (RGB, VGG-M)	37.9	72.5	56.8	79.3
Zimmermann[et al] [4] (RGB+D)	55.8	67.3	69.1	79.9
Naïve YOLO v3 (RGB+D)	58.1	79.8	72.8	86.6
Ours (RGB, S+R)	57.5	95.2	69.8	93.8
Ours (RGB+D, R)	39.0	82.1	56.7	87.3
Ours (RGB+D, S)	59.9	93.7	72.5	93.5
Ours (RGB+D, S+R)	**68.7**	**96.5**	78.6	**95.3**

TABLE III: Precision-recall curves for 3D centroids on a 60-sec sequence of our real-world test set. Solid lines correspond to an evaluation radius of 0.5m, dashed 0.25m. Crosses are at peak-F1. For our method, S stands for synthetic, R for real training data.

joints, whereas our test sequence has been annotated with 3D centroids to be fair to the baseline methods.

Methods which use a geometric 3D point cloud representation [2], [4], [6] (■, ■, ■) are more limited in recall compared to our proposed approach, which uses an image-based representation and exploits complementary RGB+D information through feature fusion. If we extend our evaluation to the (wider) RGB sensor FOV, our method makes the most out of the available data, and our peak-F1 margin over the naïve baseline increases to around +13% at $d = 0.5\text{m}$.

B. Qualitative results

In initial 2D detection experiments, we observed that almost no misdetections occur when using a pretrained, RGB-only COCO model, despite the unusual appearance of persons in our scenario. Sometimes, however, 2D bounding boxes get split in two when foreground occluder objects, such as the handle bar grip of our robot platform or a human arm, protrude into the camera FOV. This shows that exploiting depth information is important as it may help in segmenting foreground from background. It also shows that 2D bounding boxes are not an optimal representation, which provided our motivation for regressing 3D centroids end-to-end without relying on such an intermediate representation.

Qualitative 3D detection results are shown in Figures 1, 4 and 5. As in 2D, many baseline methods have problems in localizing persons under partial occlusion, and sometimes

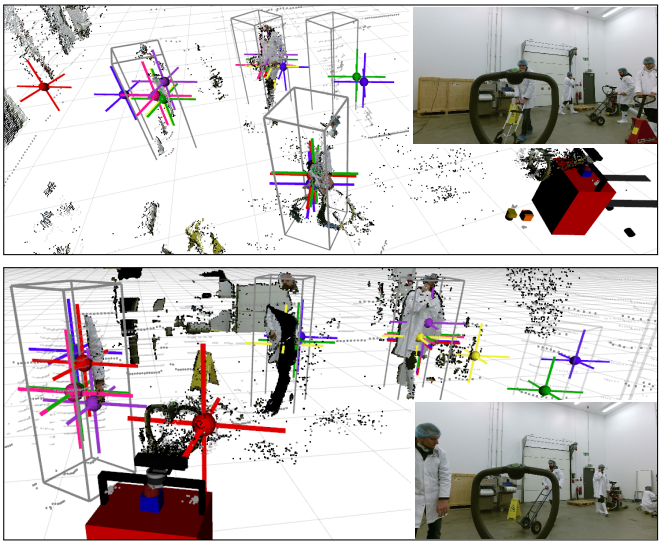


Fig. 4: Qualitative 3D detection results at peak-F1 from a scene of our RGB-D dataset. Colors from Table III; grey is groundtruth.

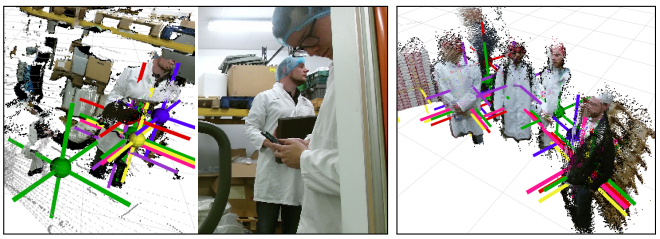


Fig. 5: Results on two further, more cluttered scenes.

even place centroids onto the scene background (such as shelves, pallets or walls). Furthermore, all baselines except for the RGB variant of [3] are strongly affected by missing depth data at the image boundaries, or at far distances. Our proposed RGB+D method is more robust in both regards.

VI. CONCLUSION

In this paper, we presented a real-time approach to the 3D human detection task, based upon the YOLO v3 architecture that we extended with an efficient RGB+D fusion scheme, 3D centroid regression, and depth-aware augmentation. Our learning strategy benefits from both synthetic and real-world training data. We demonstrated that it is possible to learn very precise 3D localization from our diverse, synthetic dataset, and showed on a subset of our challenging real-world intralogistics dataset that our method achieves higher detection accuracy than state-of-the-art baselines.

Our approach easily extends to other RGB(-D) sensors with different intrinsics, as regenerating our entire synthetic dataset with a different sensor setup only takes around 5 minutes of manual effort, while obtaining real-world 3D training labels for fine-tuning requires no manual 3D annotation.

ACKNOWLEDGMENT

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 732737 (ILIAD). We would like to thank Dennis Grießer, Michael Hernandez and Sarah Aghaie for their help with initial experiments and 3D models, as well as the ILIAD consortium for help with recording of the real-world dataset.

REFERENCES

[1] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. European Conference on Computer Vision (ECCV)*, 2016.

[2] A. Vasquez, M. Kollmitz, A. Eitel, and W. Burgard, "Deep detection of people and their mobility aids for a hospital robot," in *Proc. European Conference on Mobile Robotics (ECMR)*, 2017.

[3] M. Kollmitz, A. Eitel, A. Vasquez, and W. Burgard, "Deep 3D perception of people and their mobility aids," *Robotics and Autonomous Systems*, vol. 114, pp. 29–40, 2019.

[4] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, "3D human pose estimation in RGBD images for robotic task learning," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[5] T. Linder, D. Griesser, N. Vaskevicius, and K. Arras, "Towards accurate 3D person detection and localization from RGB-D in cluttered environments," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'18) – Workshop on Robotics for Logistics in Warehouses and Environments Shared with Humans*, 2018.

[6] M. Munaro and E. Menegatti, "Fast RGB-D people tracking for service robots," *Autonomous Robots (AURO)*, vol. 37, no. 3, pp. 227–242, 2014.

[7] O. Mees, A. Eitel, and W. Burgard, "Choosing smartly: Adaptive multimodal fusion for object detection in changing environments," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 151–156.

[8] J. Guerry, B. L. Saux, and D. Filliat, ""Look at this one": Detection sharing between modality-independent classifiers for robotic discovery of people," in *Proc. European Conference on Mobile Robotics (ECMR)*, 2017, pp. 1–6.

[9] W. Ali, S. Abdelkarim, M. Zidan, M. Zahran, and A. E. Sallab, "YOLO3D: End-to-end real-time 3D oriented object bounding box detection from LiDAR point cloud." in *European Conference on Computer Vision (ECCV) Workshops*, 2018.

[10] M. Simon, S. Milz, K. Amende, and H. Gross. (2018) Complex-YOLO: Real-time 3D object detection on point clouds.

[11] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] B. Lewandowski, J. Liebner, T. Wengefeld, S. Müller, and H. Gross, "Fast and robust 3D person detector and posture estimator for mobile robotic applications," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 4869–4875.

[13] T. Ophoff, K. Van Beeck, and T. Goedem, "Exploring RGB+Depth fusion for real-time object detection," *Sensors*, vol. 19, no. 4, 2019.

[14] D. Feng, C. Haase-Schuetz, L. Rosenbaum, H. Hertlein, F. Duffhauss, C. Glaser, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," arXiv:1902.07830, 2019.

[15] M. Gao, J. Jiang, G. Zou, V. John, and Z. Liu, "RGB-D-based object recognition using multimodal convolutional neural networks: A survey," *IEEE Access*, vol. 7, pp. 43 110–43 136, 2019.

[16] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6517–6525.

[17] R. Girshick, "Fast R-CNN," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 39, no. 6, pp. 1137–1149, 2017.

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

[20] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 23–30.

[21] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel, "Implicit 3D orientation learning for 6D object detection from RGB images," in *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[22] M. Danielczuk, M. Matl, S. Gupta, A. Li, A. Lee, J. Mahler, and K. Goldberg, "Segmenting unknown 3D objects from real depth images using Mask R-CNN trained on synthetic data," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[23] J. Shotton, A. Fitzgibbon, A. Blake, A. Kipman, M. Finocchio, B. Moore, and T. Sharp, "Real-time human pose recognition in parts from a single depth image," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[24] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[25] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Proc. European Conference on Computer Vision (ECCV)*, ser. LNCS, vol. 9906. Springer International Publishing, 2016, pp. 102–118.

[26] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[27] T. Linder, M. J. Hernandez Leon, N. Vaskevicius, and K. O. Arras, "Towards training person detectors for mobile robots using synthetically generated RGB-D data," in *Computer Vision and Pattern Recognition (CVPR) 2019 Workshop on 3D Scene Generation*, 2019.

[28] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[29] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[30] M. Fabbri, F. Lanzi, S. Calderara, A. Palazzi, R. Vezzani, and R. Cucchiara, "Learning to detect and track visible and occluded body joints in a virtual world," in *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[31] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "Synthetic occlusion augmentation for 3D human pose estimation with volumetric heatmaps," in *European Conference on Computer Vision (ECCV) Workshops*, 2018.

[32] T. Linder, S. Breuers, B. Leibe, and K. O. Arras, "On multi-modal people tracking from mobile platforms in very crowded and dynamic environments," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 5512–5519.

[33] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "Metric-scale truncation-robust heatmaps for 3D human pose estimation," in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2020.

[34] J. Guo, H. He, T. He, L. Lausen, M. Li, H. Lin, X. Shi, C. Wang, J. Xie, S. Zha, A. Zhang, H. Zhang, Z. Zhang, Z. Zhang, S. Zheng, and Y. Zhu, "GluonCV and GluonNLP: Deep learning in computer vision and natural language processing," *Journal of Machine Learning Research*, vol. 21, no. 23, pp. 1–7, 2020.

[35] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2003.

[36] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Asian Conference on Computer Vision (ACCV)*. Springer, 2016, pp. 213–228.

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Proc. European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[39] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[40] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[41] Z. Zhang, T. He, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of freebies for training object detection neural networks," arXiv:1902.04103, 2019.

[42] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. Journal of Computer Vision*, vol. 88, no. 2, 2010.