Towards Training Person Detectors for Mobile Robots using Synthetically Generated RGB-D Data

Timm Linder[1,2], Michael Johan Hernandez Leon[3], Narunas Vaskevicius[1], Kai O. Arras[1]
[1] Robert Bosch GmbH, Corporate Research
[2] University of Freiburg [3] University of Stuttgart
{timm.linder,narunas.vaskevicius,kaioliver.arras}@de.bosch.com

Abstract

We explore how we can use synthetically generated RGB-D training data from a near photo-realistic game engine to train modality-specific person detectors. We perform ablation studies on a challenging, real-world dataset which we recorded using a Kinect v2 RGB-D sensor in multiple warehouse environments. Through extensive use of domain randomization techniques, we synthesize a realistic and highly varied training set of challenging intralogistics scenarios as observed from a mobile robot, comprising persons in confined and cluttered indoor spaces. We then train the detector layers of a YOLOv3 model from scratch on our synthetic RGB and jet-encoded depth images. While for the RGB case, we still observe a domain gap of 6 points in mAP compared to a pretrained COCO model, results indicate that by exploiting simulation, an immense manual labeling effort needed to prepare large-scale datasets such as MS COCO might be unnecessary for the depth modality. We further find that filtering of highly occluded groundtruth bounding boxes during training, as well as modeling of time-of-flight sensor noise characteristics has a positive impact on model performance. We also provide an initial set of qualitative results on our real-world dataset.

1. Introduction

Robust detection and tracking of persons in real-time from an ego-perspective is important for robots to operate safely and efficiently in human environments. One challenge encountered when training object detectors for robotics applications is that the sensor setup can be multi-modal, and vary significantly between robots. Furthermore, the particular application domain may provide additional challenges which are underrepresented in commonly utilized object detection datasets, such as persons wearing protective clothing and thus being of similar appearance.

Figure 1: Using Unreal Engine 4, we synthetically generate crowded RGB-D sequences to train robust person detectors. We evaluate performance of the resulting YOLOv3 models on a challenging intralogistics dataset which we recorded. In the top image, red contours denote small groundtruth objects that we filtered out during training for better results.

In this work, we want to examine if we can replace large-scale annotated object detection datasets such as MS COCO with synthetically generated data from game engines, in our case Unreal Engine 4. In particular, we are interested in synthesizing RGB-D sensor data, as no COCO-scale object or person detection datasets exist in RGB-D. As a step towards training a joint RGB-D detector from scratch, we attempt to train modality-specific person detectors from synthetic data to gain further insights into which aspects are particularly important for the modality at hand.

The contributions of this work are the following: (i) We synthesize a realistic dataset for challenging intralogistics scenarios with persons in confined and cluttered spaces, observed by a robot with an RGB-D sensor. (ii) We train a

real-time capable single-shot detector (YOLOv3) with our synthetic data and evaluate it on a real RGB-D dataset consisting of 3,100 labeled frames, which we acquired in three different warehouse environments using two different AGV platforms. The preliminary results show that by using synthetic images, we may avoid the huge labeling effort needed to prepare large-scale datasets like MS COCO, especially for the depth modality. (iii) To the best of our knowledge, we are the first to attempt to train a person detector on synthetically generated depth data. (iv) Furthermore, we perform an ablation study to investigate how the different simulation aspects influence the performance of the trained model on real-world data.

2. Related work

Large-scale annotated datasets such as ImageNet [22] and MS COCO [12] enabled a rapid advancement and benchmarking of deep learning methods in the RGB domain. Examples of benchmarks specific to person detection in RGB images from urban environments include Caltech [3] and CityPersons [27, 2]. However, for robotic applications usually full 3D awareness of surroundings including humans is required [13]. Therefore, additional sensing modalities such as depth are used to facilitate real-time processing.

In the robotics community, various RGB-D datasets have been introduced for person detection covering different depth sensing technologies such as structured light, time-of-flight, or stereo [23], [9]. Existing available Kinect v2 datasets for person detection have been recorded in indoor and outdoor environments at a university campus [1, 18, 17, 26] or a hospital [10]. Some lack person annotations when depth is not available [17], or are destined for multi-class detection involving persons with walking aids [10]. In addition, all these datasets are an order of a magnitude smaller and show much less variation in comparison to the datasets available for the RGB modality.

Besides transfer learning [8] and other weakly supervised approaches, one of the means to deal with the lack of annotated data is simulation. Synthetic data generation with game engines has been explored in autonomous driving settings [21], [20]. Although the corresponding datasets include person annotations, they are focused on outdoor urban environments. A large-scale synthetic dataset based on realistic human models [14] was introduced in [24]. The dataset is focused on human pose estimation and contains one person per scene. Also, there seems to be a lack of 3D and physical awareness in the scenes. A photorealistic dataset for multi-person pose estimation and tracking was generated using a game engine in [5]. It covers urban indoor and outdoor environments, different lighting conditions, crowded scenes, occlusions and variety of viewpoints. However, it does not focus on the egocentric viewpoint typical to robotic applications and more importantly does not provide depth images.

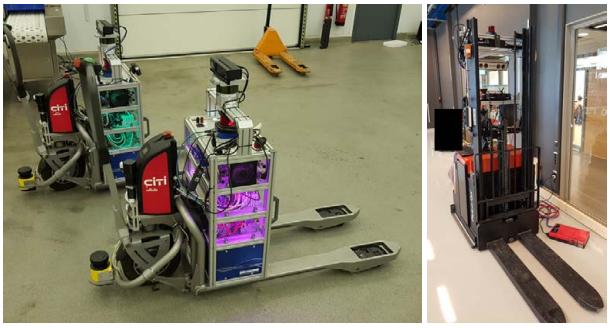


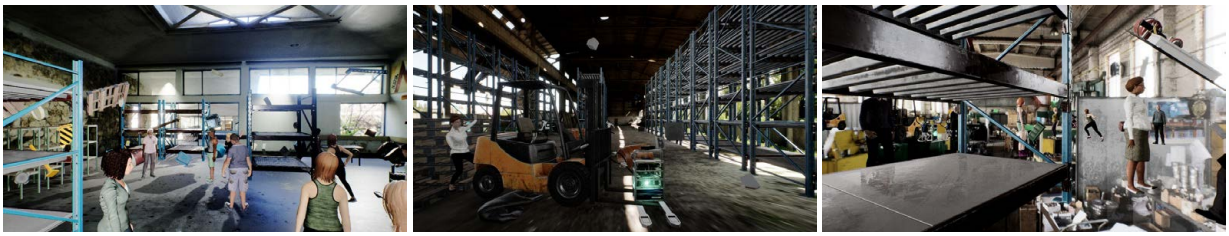
Figure 2: Two types of autonomous guided vehicles which were used to record the real-world RGB-D intralogistics dataset (with a Kinect v2) for validation and testing.

Synthetic RGB-D data generation for a variety of scene understanding tasks, including object detection, has been explored in [16, 15, 6]. However, in contrast to our work, they focus on static environments and rigid objects, and not humans with their large variation in shape and appearance.

3. Dataset

Real-world intralogistics dataset As a basis for our experiments with synthetic data, we use a real-world RGB-D dataset for validation and testing, recorded using two autonomous forklifts with different sensor setups, which we initially described in [13]. We significantly extended the number of frames labeled with 2D person bounding boxes to around 3.1k images total (1.5k train + 0.5k validation + 1.1k test), spanning several days at four different locations (two warehouses, a small food factory, and a robotics laboratory with forklifts and warehouse shelves). The real-world dataset contains both sequences with very few people, and very crowded scenes with up to around 20 people, sometimes all in very similar clothing (Fig. 1 bottom). This makes it difficult for standard RGB detectors to discern and properly localize individual persons, such that proper use of depth information could be highly beneficial.

Synthetic dataset from UE4 For our experiments, we have built a synthetic RGB-D dataset using Unreal Engine 4, currently consisting of six scenes and 15k frames in total. From each scene, we generate 2.5k RGB-D image pairs with corresponding instance segmentation masks, which are used to compute groundtruth bounding boxes. Scenes 1-3 (Fig. 3a) are custom-made and contain various warehouse shelves and objects. The background of these scenes is randomized at regular intervals from 25 publicly available HDR images.



(a) Scenes 1–3

(b) Scenes 4–6

Figure 3: Our synthetic training set comprises 6 different scenes. The first three contain real 2D HDR images as backgrounds, whereas the latter three are completely composed of 3D objects. Lighting, foreground 3D objects, etc. are randomized.

Scenes 4-6 (Fig. 3b) are based upon publicly available environments representing a warehouse, a train station, and an outdoor factory environment. All of the scenes feature randomization of light sources in terms of their placement and intensities.

3D augmentation We enrich all scenes with random flying 3D objects (by using the UE4 physics engine to prevent objects from overlapping), moving forklifts and pallet trucks.

Synthetic humans We use a set of 24 person meshes that were generated synthetically using Adobe Fuse. Around 50 different animations, including idling, 8 walking styles, dancing and jumping, are applied randomly. Using the existing material masks, we randomly augment the clothing colors and texture.

Character navigation and camera movement For moving the human characters, as well as the robot platform with the RGB-D sensor, through the scene, we rely on UE4's navigation mesh system to perform pathfinding in walkable areas. Simplified collision meshes are used to prevent colliding with the environment and other characters.

Sensor modeling The RGB and depth imagers are each simulated by a separate virtual camera, with resolution of 1920×1080 and 512×424 px, respectively. Via back-projection, we simulate the ∼5 cm offset between IR emitter and receiver, leading to shadowing effects especially at close-range (for instance visible in Fig. 4c). Based upon experimental measurements from [11] and comparison with our real-world data, we also empirically model several acquisition-based errors such as depth and amplitude distortion, axial noise, illumination interference, and material IR responses (which is particularly important *e.g.* for dark clothing, and approximated via PBR material diffuse and roughness values). More complex effects such as flying pixels or multi-path interference are currently not implemented, and would require *e.g.* raytracing techniques.

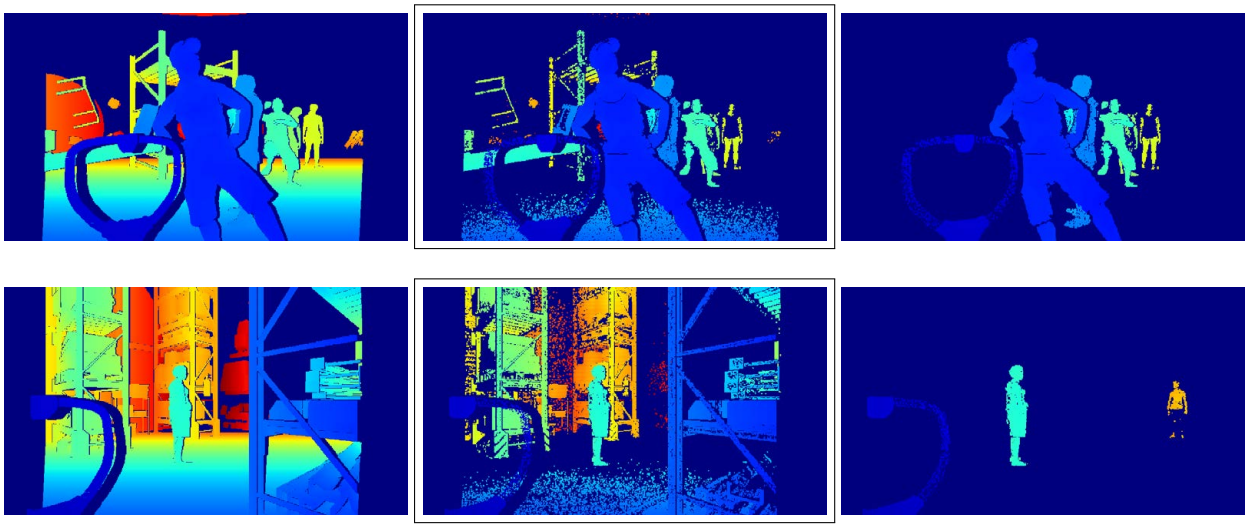
The depth sensor's horizontal field of view is smaller than the field of view of the color sensor. Therefore, we filter and adjust the groundtruth bounding boxes in post-processing such that depth groundtruth boxes do not extend beyond the field of view of the depth sensor. Registration of RGB and depth images is performed, like for the real-world dataset, using `iai_kinect2` [25] while assuming a perfect intrinsic and extrinsic calibration.

4. Experiments

We use the MxNet implementation of the YOLOv3 detector [19] (with DarkNet53 pretrained on ImageNet) for our experiments, which showed the best compromise between speed and accuracy during our initial experiments on real-world data, and outperformed the original implementation. We train YOLOv3-416 on single V100 (32 GB) GPU at a batch size of 64. Standard 2D image augmentation techniques are used. For training from scratch, we train with a single output class (person) and use a learning rate schedule with 5 initial warmup epochs leading to a learning rate of 0.001, which is decayed by a factor of 0.1 at epochs 160 and 180. We disable input shape randomization for faster training at a larger batch size. For fine-tuning over 40 epochs, an initial learning rate of 0.0001 is decayed by a factor of 0.3 after 20 epochs.

(a) No shading / Default / No static 3D objects

(b) No shading / Default / Strong post-processing effects



(c) No Kinect v2 sensor noise model / Default / No static 3D objects

Figure 4: Example synthetic RGB and depth images showing effects examined in ablation studies. Our proposed configuration, which yielded the best results in our experiments, is depicted in the center of each row.

To train on depth images, we use a jet color map encoding as proposed by Eitel et al. [4] since it is faster than other encodings such as HHA encoding [7] which rely on expensive surface normal computations. Also, it allows for transfer learning by applying pretrained ImageNet RGB weights to depth data. An alternative, besides also training the feature extractor on synthetic data, might be cross-modal distillation [8] for training directly on 1-channel depth images.

Ablation studies We perform a series of ablation studies, where we selectively disable or enable certain effects (shading, static 3D objects, stronger post-processing effects in RGB / sensor noise modeling, static 3D objects in depth) as visualized in Figure 4. All these variations of the dataset are generated in parallel (by pausing the simulation), in order to ensure deterministic behavior and otherwise identical content. We also conduct trainings with only 7.5k instead of 15k training images, where we either uniformly subsample all frames, or selectively only use scenes 1-3 with random HDRI backgrounds, or only scenes 4-6 which are entirely modeled in 3D.

Finally, we also examine if filtering the generated synthetic groundtruth bounding boxes (by minimum area in pixels, in order to omit very tiny or highly occluded boxes) can improve training performance. Otherwise, groundtruth boxes

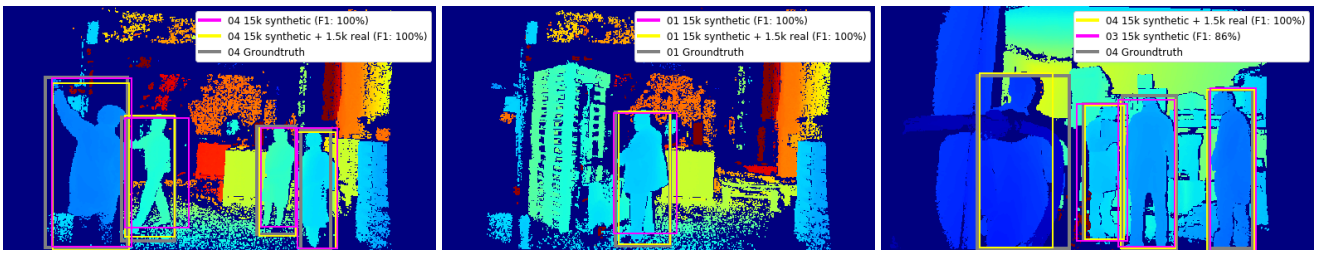


Figure 5: Qualitative results for depth-based person detection on the real-world test set. Groundtruth in gray, YOLOv3 detector trained on 15k synthetic depth images in magenta, and in combination with 1.5k real training images in yellow.

may be generated for image regions which a human annotator would never label as 'person', leading to a different distribution of bounding boxes in synthetic training and real-world test/validation data.

5. Results

Quantitative results for training with synthetic RGB and depth data, respectively, are visualized in Fig. 6 and Fig. 7. In each figure, the top diagram compares different combinations of synthetic and real data when either fine-tuning a COCO-based model, or training all detector layers from scratch. The bottom diagram shows how performance degrades (from the proposed default configuration) if certain effects are selectively disabled. Qualitative results for the depth-based detector are shown in Fig. 5.

Overall, detection performance on RGB is much better than on depth, while groundtruth bounding boxes for the latter have already been filtered to accommodate for the limited range and depth sensor's field of view. When training the detector layers from scratch, adding synthetic data clearly boosts the performance (+10.3% mAP on RGB, +6.5% on depth), compared to just training on the available 1.5k real RGB-D frames. One reason for the domain gap when using only synthetic depth data could be the lack of flying pixels in our simulation, which appear quite prominently in real-world data.

Combining both our synthetic and real-world depth data leads to detection results which are quite close to, or better than, a COCO model fine-tuned on real data (-0.4% mAP on test, +1.7% mAP on validation set; qualitatively, mainly bounding box localization seems to improve). This means that we essentially do not need MS COCO pretraining anymore for learning a depth-based detector.

However, a model pretrained on MS COCO still shows better performance on RGB, which could be either due to lack of realism, or lack of diversity in our synthetic RGB data. On RGB, the number of training images (15k vs. 7.5k) seems to have a much larger impact (-5.2%) than on depth (-1.3%). Instead, for depth, the absence of static 3D objects seems to have a stronger impact (-3.3% vs. -2.0% on RGB).

Also, in our experiments on the RGB modality, we find that filtering of synthetic groundtruth bounding boxes based upon a minimum amount of visible pixels in the instance mask (Fig. 1, top), can improve mAP by approx. +2%.

The performance gain by modeling ToF depth sensor noise is somewhat limited (+1.2% on test / +0.2% on validation). However, not all noise sources have been modeled, and registration shadowing (IR emitter offset) was always enabled.

6. Conclusion

The main focus of our work is on person detection in specialized use-cases, characterized by different sensory setups (e.g. multi-modal, first-person perspective) and a lack of relevant training data. Available datasets poorly cover such use-cases in relevant quantity and variation. Thus, we took initial steps towards exploring the use of synthetic data from simulation (of environments, objects of interest and sensors), combined them with a small real-world dataset, and performed ablation studies to explore the impact of different data combinations, rendering effects and sensor models as well as static object configurations. Specifically, we evaluated this approach for the problem of RGB-D person detection in professional environments from a mobile robot platform. What makes this additionally a challenge is that in the considered warehouse and factory environments, people often look alike due to their work-wear and are occluded frequently.

Our results indicate that training depth-based person detectors from synthetic data can achieve performance comparable to, or even better than, transfer learning from a COCO-based model.

This is an important step towards training a *joint* RGB-D detector from synthetic data (which we purposefully did not do so far to allow for studying the characteristics of the RGB and depth modalities independently). ImageNet, MS COCO and other common large-scale datasets do not provide any depth data, allowing only for transfer learning

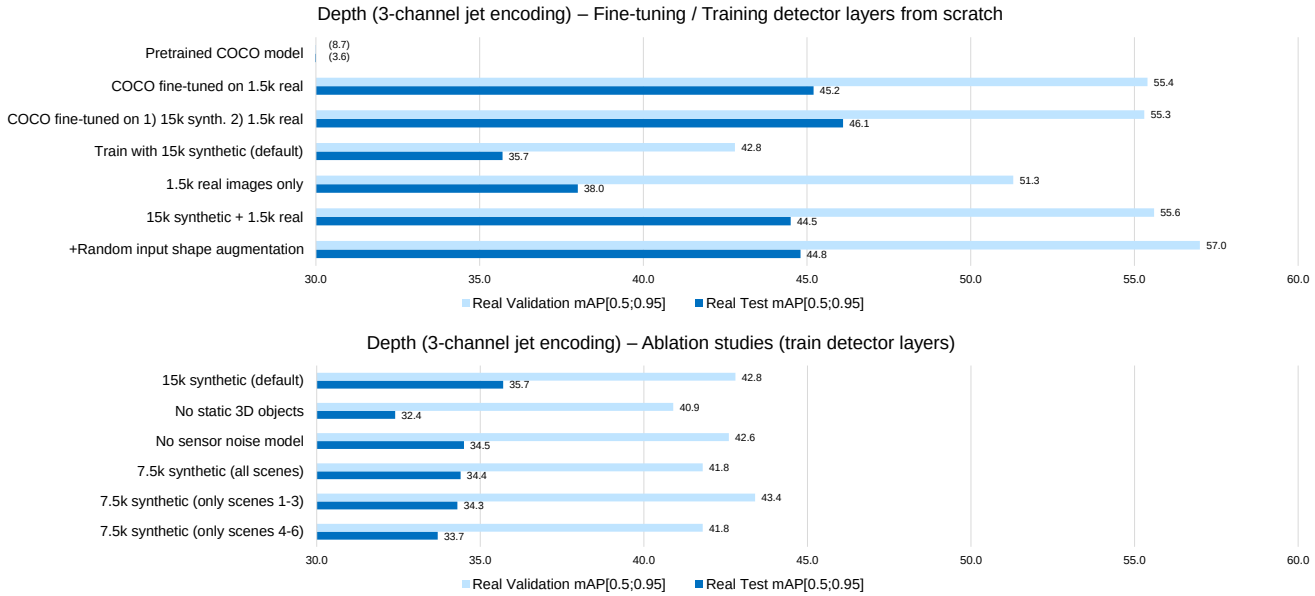


Figure 6: Synthetic Depth training results. Validation and testing on real-world intralogistics dataset.

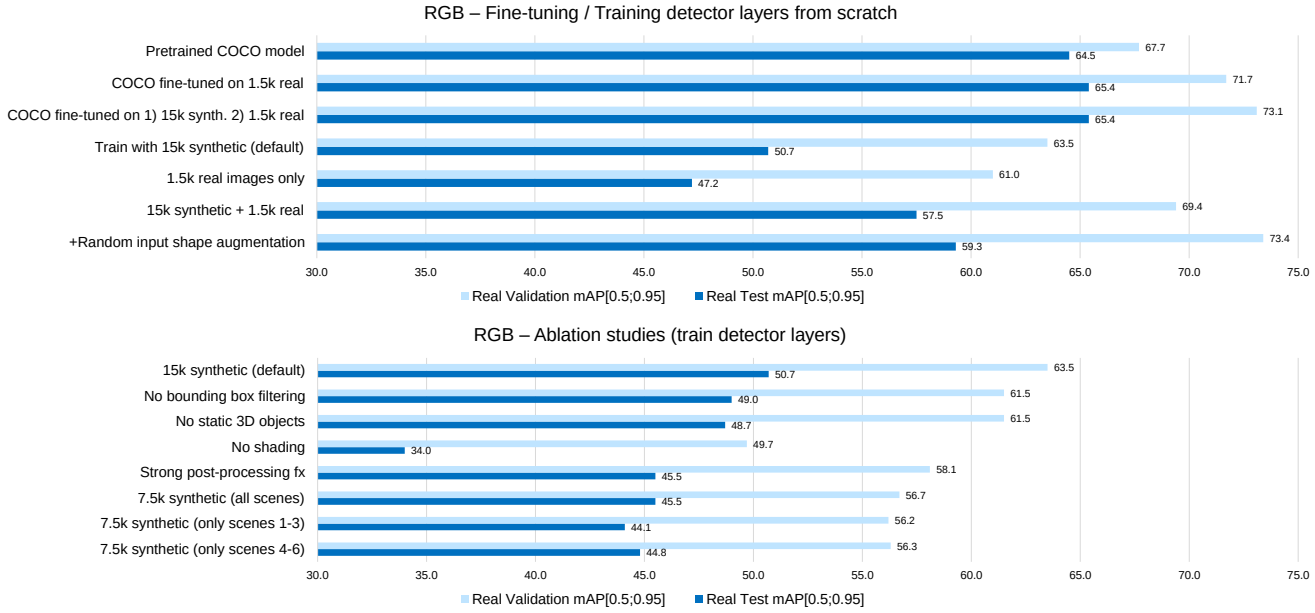


Figure 7: Synthetic RGB training results. Validation and testing on real-world intralogistics dataset.

methods to be applied that rely on reusing RGB information and do not capture the qualitative differences between the two sensor modalities. Instead, we explicitly want to exploit complementary depth information where RGB alone may not suffice, *e.g.* when discriminating partially occluded persons of very similar appearance.

Future work We also want to increase the overall size of our synthetic dataset, and experiment with the variety and type of human models, incorporating also photogrammetrically derived (3D-scanned) meshes. It could also be interesting to study the effect of clothing animation on RGB-D detection performance. It would make sense to also learn the feature extractor from synthetic data, which has been shown in [16] to yield results superior to ImageNet pretraining for RGB-D indoor room segmentation. Finally, extending the groundtruth of our dataset to cover additional tasks, such as articulated 3D human pose estimation from RGB-D [28] (where manual annotation is very challenging), would not pose much extra effort.

Acknowledgements This work has received funding from the EU H2020 research and innovation programme under grant agreement No 732737 (ILIAD). We would like to thank Sarah Aghaie and our ILIAD project partners for their help with the human meshes and recording of the real-world dataset.

References

[1] T. Bagautdinov, F. Fleuret, and P. Fua. Probability occupancy maps for occluded depth images. In *CVPR*, 2015.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[3] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 34, 2012.

[4] A. Eitel, J.T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard. Multimodal deep learning for robust RGB-D object recognition. In *IROS*, 2015.

[5] 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 *ECCV*, 2018.

[6] G. Georgakis, A. Mousavian, A. C. Berg, and J. Kosecka. Synthesizing training data for object detection in indoor scenes. In *RSS*, 2017.

[7] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik. Learning rich features from RGB-D images for object detection and segmentation. In *ECCV*. 2014.

[8] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. arXiv:1507.00448v2.

[9] R. Hanten, P. Kuhlmann, S. Otte, and A. Zell. Robust real-time 3D person detection for indoor and outdoor applications. In *ICRA*, 2018.

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

[11] E. Lachat, H. Macher, T. Landes, and P. Grussenmeyer. Assessment and Calibration of a RGB-D Camera (Kinect v2 Sensor) Towards a Potential Use for Close-Range 3D Modeling. *Remote Sensing*, 7, 2015.

[12] 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 *ECCV*, 2014.

[13] T. Linder, D. Griesser, N. Vaskevicius, and K.O. Arras. Towards accurate 3D person detection and localization from RGB-D in cluttered environments. In *IROS Workshop*, 2018.

[14] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M.J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics*, 34(6), 2015.

[15] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison. SceneNet RGB-D: 5M photorealistic images of synthetic indoor trajectories with ground truth. arXiv:1612.05079, 2016.

[16] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison. SceneNet RGB-D: Can 5M synthetic images beat generic imagenet pre-training on indoor segmentation? In *ICCV*, 2017.

[17] O. Mees, A. Eitel, and W. Burgard. Choosing smartly: Adaptive multimodal fusion for object detection in changing environments. In *IROS*, 2016.

[18] T. Ophoff, K. Van Beeck, and T. Goedem. Exploring RGB+Depth fusion for real-time object detection. *Sensors*, 19(4), 2019.

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

[20] S.R. Richter, Z. Hayder, and V. Koltun. Playing for benchmarks. In *ICCV*, 2017.

[21] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez. The SYNTHIA dataset: A large collection of synthetic images for sem. seg. of urban scenes. In *CVPR*, 2016.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3), 2015.

[23] L. Spinello and K. O. Arras. People detection in RGB-D data. In *IROS*, San Francisco, USA, 2011.

[24] G. Varol, J. Romero, X. Martin, N. Mahmood, M.J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017.

[25] T. Wiedemeyer. IAI Kinect2. `https://github.com/code-iai/iai_kinect2`, 2014 – 2015.

[26] N. Wojke, R. Memmesheimer, and D. Paulus. Joint operator detection and tracking for person following from mobile platforms. In *FUSION*, 2017.

[27] S. Zhang, R. Benenson, and B. Schiele. CityPersons: A diverse dataset for pedestrian detection. In *CVPR*, 2017.

[28] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox. 3D human pose estimation in RGBD images for robotic task learning. In *ICRA*, 2018.

Towards Training Person Detectors from Synthetic RGB-D Data

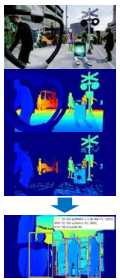
Timm Linder, Narunas Vaskevicius, Michael J. Hernandez Leon, Kai O. Arras

BOSCH Robotics Research · UNI FREIBURG · ILIAD · Universität Stuttgart

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

AT A GLANCE

In this workshop paper, we present initial findings on training person detectors for robots from synthetic RGB-D data using various domain randomization / augmentation techniques.

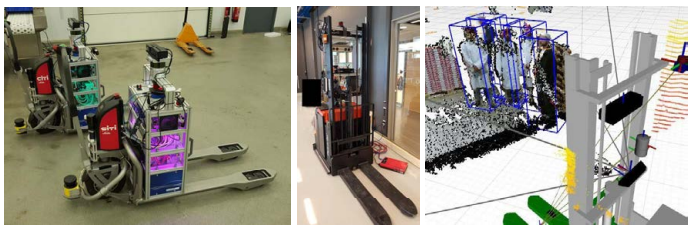


1. We **synthesize** a **diverse RGB-D dataset** using Unreal Engine 4, with persons in **cluttered + confined spaces**
2. We **train modality-specific** (depth / RGB) single-shot person detectors from synthetic data, and **evaluate on real-world data** acquired by AGVs in multiple intralogistics environments
3. We are the first (to our best knowledge) to **report on training depth-based** multi-person detectors solely **from synthetic data**
4. Several **ablation studies**, e.g. Kinect v2 depth noise modelling

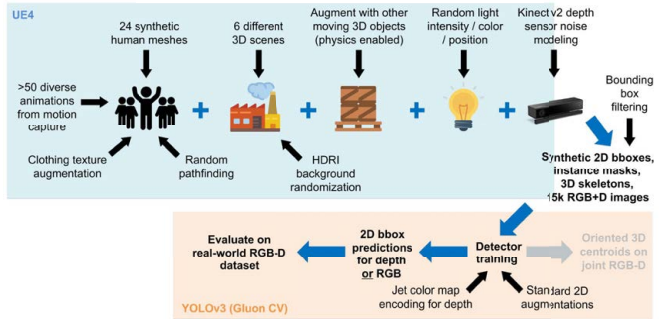
MOTIVATION

Can we rely mostly on synthetic data for training RGB-D person detectors?

- RGB-D important for robotics because **3D localization of objects** is essential
- Leverage **complementary depth information** where **RGB fails** (bad lighting, reflections and pictures of humans on walls → lack of 3D geometry, etc.)
- Most existing papers/datasets in RGB-D focus on static scenes (e.g. indoor segmentation), not **humans with highly varying shape and appearance**
- **No COCO- or ImageNet-scale datasets in RGB-D** for person detection with sufficient **scene diversity** (due to **lack of crowdsourcing**)
- **Manual labelling** for some 3D tasks is very difficult
- Initial focus on **2D bbox detection** (later 3D centroids/bboxes + 3D body joints)



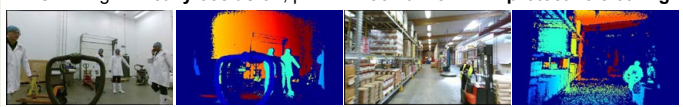
SYNTHETIC DATASET + METHOD



- We **model** depth distortion, reg. offset, lateral/axial noise, material IR response
- We train all **YOLOv3 detector layers from scratch** using synthetic data
- Currently still relying on **ImageNet pre-training** for feature extractor

REAL DATASET FOR TESTING + FINE-TUNING

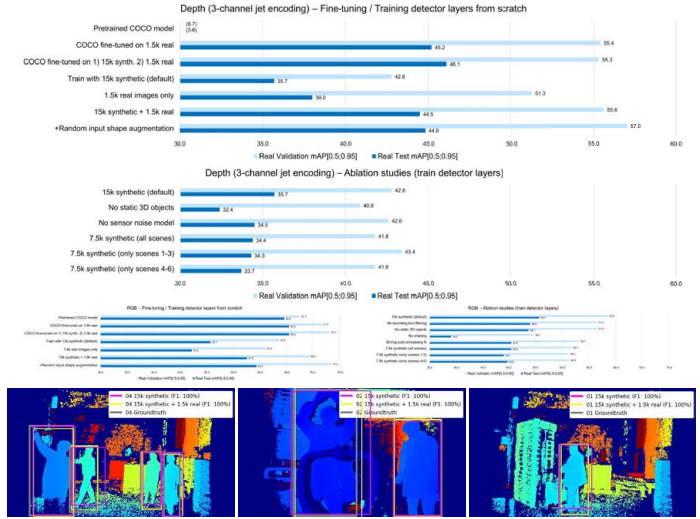
- Recorded in **3 different warehouse environments** by 2 different AGVs
- Manually labelled **3,100 RGB-D frames** with 2D person bounding boxes
- **1,100 test images** + 1,500 for fine-tuning + 500 for validation
- Challenges: **Heavy occlusion**, persons **look alike** due to **protective clothing**



EXPERIMENTAL RESULTS

Key insights – tested on our real-world RGB-D dataset:

- **Qualitative results look promising** when training just on synthetic data
- **Adding synthetic data** (15k frames) to limited available real training data (1.5k frames) **clearly boosts performance** (mAP +10.3% on RGB, +6.5% on depth)
- **On depth**: Using synthetic data + only 1.5k real images, we almost reach performance of transfer-learned COCO model → may not need COCO anymore! This could become useful for **joint end-to-end training on RGB-D**
- **Depth sensor noise modelling** leads to small improvement (+1.2%)
- **But**: Still **domain gap on RGB** → add more human models, more background augmentation, more scenes, try style transfer using GANs
- **Filtering of groundtruth bounding boxes** based upon distance and visible area is more important (shown on RGB: +2%)
- Excessive **visual effects** (blur, lens flares, etc.) hurt model performance
- **Combination of synthetic + real data** (10%) overall gives best results



CVPR 2019 Workshop on 3D Scene Generation

This work has received funding from the EU H2020 research and innovation programme under grant agreement no. 732737 (ILIAD).