5th Conference on Production Systems and Logistics

Investigation of Deep Learning Datasets for Warehousing Logistics

Dimitrij-Marian Holm[1], Philipp Junge[1], Jérôme Rutinowski[2], Johannes Fottner[1]
[1]*Technical University of Munich, Chair of Materials Handling, Material Flow, Logistics, Garching b. München, Germany*
[2]*TU Dortmund University, Chair of Material Handling and Warehousing, Dortmund, Germany*

Abstract

Deep Learning for Computer Vision holds great potential in warehousing logistics, for example for applications such as mobile robots or autonomous forklifts. However, the availability of labelled image datasets within this area is limited. To address this problem, we benchmarked two different datasets, LOCO (Logistics Objects in Context) and TOMIE (Tracking Of Multiple Industrial Entities), to find out, if these datasets can be used interchangeably. Therefore, we examine the usability of these datasets for Object Detection tasks using the YOLOv7 framework. For this we trained several networks and compared them with each other. A deep analysis between these two datasets shows that they are quite different and only suitable for specific tasks which are not interchangeable, despite having emerged from the same research domain. More thorough investigations are performed to find the reasons for this lack of compatibility. To close the gap between LOCO and TOMIE, a synthetic data generation pipeline for pallets is developed and 18,000 synthetic pallet images are rendered. Furthermore, models are trained based on the synthetic data and compared with the models trained on real data. The synthetic data generation pipeline successfully closes the reality gap, and the performance on TOMIE is increased, but the performance on LOCO remains significantly weaker, in comparison. To develop a deeper understanding of this behaviour we examine the underlying datasets and the reasons for the performance difference are identified.

Keywords

Warehousing Logistics; Datasets Generation; Deep Learning; Object Detection

1. Introduction

Robotic systems, such as AGVs (autonomous guided vehicles) and autonomous forklift trucks are encountered more and more in warehousing environments, solving tasks, like recognising, storing and collecting pallets and other objects of interest [1]. To enhance the abilities of such robotic systems, they require the capability of a deeper understanding of the environment by using sensors such as cameras and algorithms to extract semantic information. For this task, there still remains a huge untapped potential for Deep Learning (DL) and Computer Vision (CV). For Deep Learning however, lots of data is necessary, which is sparsely available for warehousing environments [2].

Since existing datasets from the context of warehousing logistics are limited, it first leads us to LOCO (Logistics Objects in COntext), which was the first dataset of its kind [3]. Also, TU Dortmund University recently published a tracking dataset in the field of warehousing logistics called TOMIE (Tracking Of Multiple Industrial Entities) [4].

DOI: https://doi.org/10.15488/15311
ISSN: 2701-6277

Both datasets are designed for Object Detection in warehousing environments, providing semantic data of logistical entities within their images. In addition, we address the question to what extent synthetic data can improve the combination of both datasets in terms of Object Detection results.

Therefore, LOCO as well as TOMIE are being investigated in terms of their usability for DL. The use of synthetic data is also evaluated, and an image generation pipeline is developed for this purpose. Experiments are conducted using the state of the art Object Detection framework YOLOv7 [5].

2. Related Work

Object Detectors are normally used on a special domain, therefore the need for a specialised dataset is present. In the field of warehousing logistics, there was no publicly available dataset until the release of LOCO [3] as well as the recently published paper of TOMIE [4]. Nevertheless, there is some work that dealt with Object Detection in logistics using real data: In [6] a survey on DL-based Object Detection in industrial manufacturing lines was conducted. Another paper [1] deals with the automated detection of pallets by unmanned forklifts. 4,620 photos from real warehouses were used for training with the Single-Shot-Detection (SSD) architecture. The work of [7] aimed to detect pallets and their associated pallet pockets using Object Detection. For this purpose, Faster-RCNN, SSD and YOLOv4 were compared, with the result that Faster-RCNN and SSD achieve better performance, but small objects are detected significantly better by YOLOv4. Pallet detection with YOLOv5 was investigated in [8]. 1,350 images were captured with three different cameras at different times of the day. In addition to pallet Object Detection with SSD, [9] used depth data to extract the 3D pallet point-cloud model for accurate positioning. For this purpose, more than 1,000 images of pallets were taken under different lighting conditions and at normal forklift reach.

Due to the lack of datasets in the real logistics environment, there are many efforts to generate them synthetically. There are different ways of creating synthetic data, one of which is rendering images with 3D software, e. g. with Blender or Unity. This can be roughly divided into trying to render photo-realistic images, imitate real world parameters as close as possible, or using Domain Randomisation (DR), for which the realism is not that important. DR is an approach where parameters of the source domain are randomized with the idea that the target domain is recognised as just another variation of the source domain by the model [2]. [10] focused on the detection of retail-objects, using a DR approach with random 3D objects in the background of the objects of interest. The Hamburg University of Technology [11] dealt with the pose estimation of a Euro-pallet, also using Blender. However, the focus here lays more on photorealism, the textures consist of RGB images from a real camera. An mAP of 0.94 was reached on own test data. In [2], an industry-based synthetic dataset consisting of small load carriers was used to evaluate their data generation pipeline using Blender and DR. In addition to rendering images and compositing real data, there is another promising way to generate synthetic data, which is generative AI. AI has been given a whole new meaning by Large Language Models like ChatGPT, but are also potentially useful in the field of Object Detection. GANs (Generative Adversarial Networks) for instance, can be helpful in domain adaptation in order to bring synthetic images closer to reality [12]. However, research on generative AI for Object Detection in warehousing logistics is still at the beginning.

In the context of warehousing, only one synthetic dataset is known to us [13]. This dataset, however, is suited for re-identification and not for Object Detection tasks. One further dataset, which is currently available only for collaborators and partners, is SORDI [14], created to tackle the lack of industrial synthetic datasets. 200,000 bounding box annotated images were rendered, containing eight different assets in 32 scenarios, resulting in more than 1 million bounding boxes.

3. Methodology

The goal of this paper is the investigation of the LOCO and TOMIE dataset in terms of usage potential for Object Detection. We evaluated if the Object Detection performance of LOCO can be improved by using a modern framework, the TOMIE dataset or self-created synthetic data. The best-case scenario would be a model which generalises well on all data in context of logistics by using all three sources of data.

First, data exploration and comparison for both LOCO and TOMIE is conducted. With the base models of both real datasets, reciprocal inference is conducted to evaluate how well each model is performing on the other dataset. Regarding synthetic data (from now on referred to as the SYNTH dataset), first, a pipeline was created using Blender, generating images of the pallet class. To bridge the reality gap, DR as well as domain knowledge are used. In this case, the latter means that we already know where and under which conditions (i. e., loaded, on shelves) pallets are often seen. After generating these images, the pipeline also must prepare the data for training and data exploration. Following that, training with YOLOv7 is conducted, leading to a base model trained on SYNTH. As with the real data, inference is conducted with TOMIE and LOCO to determine how well purely synthetic pallet data works in the logistics domain. With the then available base models, fine-tuning is done - for the SYNTH base model with LOCO and TOMIE, for the real datasets only with the respective other dataset. Afterwards, inference is conducted again with LOCO and TOMIE to evaluate how fine-tuning helps to generalise or if catastrophic forgetting occurs. To compare the performance of the SYNTH dataset, base models are again trained with LOCO and TOMIE also using only the pallet class. The overall approach can be seen in Figure 1.

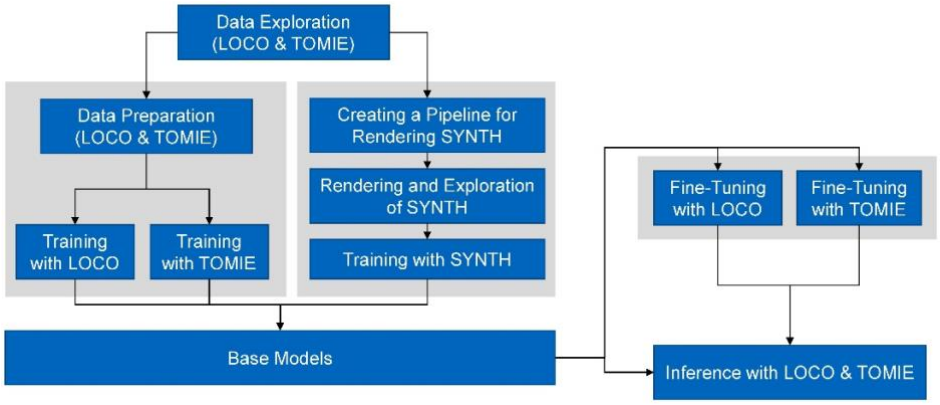

Figure 1: Training and test approach.

4. Data Exploration

4.1.1 LOCO

LOCO is a dataset which is split into five subsets, representing different warehousing logistics environments. A total of 64,993 images were captured, images 39,101 remained after removing blurred and similar images. From those, 5,593 were selected for bounding box annotation for five classes: pallets, small load carriers, stillages, forklifts and pallet trucks. In total there are 151,428 instances of those classes, which have a unbalanced class distribution. Pallets make up the biggest part of the annotations, followed by small load carriers. Forklifts occur only 598 times in the dataset. There are 496 images without any annotations, the majority of the images hold at least ten annotations and a significant number even more than 50 annotations. Also, the relative size of the objects is smaller in LOCO than in common datasets like COCO. 90% of the annotations have a bounding box size smaller than 2% of the image size. In COCO, that is only the case for 70% of the objects. The majority of LOCO annotations are even smaller than 1% of the image size.

4.1.2 TOMIE

TOMIE [4] was recorded in a research facility representing a warehousing environment. In this environment, different warehousing scenarios were recorded using six cameras in different locations, creating six different data subsets. The dataset consists of 112,860 frames and 640,936 entity instances of nine classes including pallets, stillages, small load carriers and forklifts. Compared to LOCO this means that it offers the same object classes, except for the pallet trucks. The number of annotations per class is more balanced in TOMIE than in LOCO. Nevertheless, the pallet also dominates here, with over twice as many instances as the small load carriers. Stillages are the rarest while there are still twice as many forklifts. TOMIE has few instances per image, especially compared to LOCO, the majority of the images has no more than 10 annotations. The size of the annotations is more irregular compared to LOCO. This is because the camera distance to the respective objects remains virtually the same throughout the recordings. Furthermore, there are no annotations with a bounding box smaller than 0.1% of the image size.

4.1.3 SYNTH

Due to the scope of this paper, only one of the potential classes is considered. Pallets were chosen since they play the most elementary role and occur the most in the existing datasets. Thus, the performance of the synthetic data can be measured best.

To ensure data diversity, 13 different pallet and three different warehouse assets are used for the generation of SYNTH. The pallets are randomly spawned following various parameters which are randomly selected based on a self-defined range that can be changed before rendering. This includes the pose, texture, form and quantity. In addition to labelling full pallets, 2D front and side views of pallets were added to SYNTH. Camera position, orientation, and field of view were also randomized. Distractor objects with random attributes were placed in the foreground and background. The background was randomized by changing floor and wall textures. 3D assets from the logistics context were placed in the background. Random lighting was applied, including the use of a shadow thrower to simulate poorly illuminated pallets. Rendering quality was varied, including resolution and number of samples. After rendering, some sanity post-processing was applied, like excluding pallets which are barely visible in the image, either because they are placed at the image edges, or because they are occluded.

Figure 2: Examples of the different subsets of SYNTH.

About 18,000 images were rendered. with a total of 125,937 instances, divided into 6 subsets, all with slightly different parameters. Examples for images from the dataset can be seen in Figure 2. The annotations per image distribution is located between LOCO and TOMIE with the maximum of images having 1 to 5 pallets and no images with more than 50 pallets, with seven on average. The size of the bounding boxes is quite similar to LOCO. The composition of the dataset and the split into different subsets is shown in Table 1.

Table 1: Statistics of the SYNTH dataset.

Subset	Number of Images	Number of Pallets	Notes
α	2,992	25,279	Full randomization
β	3,017	15,926	Pallets only on floor level
γ	2,989	18,797	Full randomization
δ	3,006	15,109	Pallets only on floor level
ϵ	2,991	32,544	No texture change for floor and wall
ζ	3,035	19,625	No texture change for floor and wall More pallet front side instances

5. Experiments

5.1 Real Data

For the experiments with real data, the LOCO subsets (1-5) were split into training and validation sets, i. e. 2, 3 and 5 as training sets and subsets 1 and 4 as validation sets. Using TOMIE, we split the subsets (A - F) so that the training set contains A, B, E and the sets C and D as the validation set. We used an image resolution of 1120px, a batch size of 4, an IoU of 0.5 and a confidence of 0.001. For training, the standard YOLOv7 P5 parameters and pre-trained COCO weights were used.

5.1.1 LOCO Base Model

We trained the LOCO base model with Adam, a LR (Learning Rate) of 0.001 and 5 frozen layers. Inference was conducted with LOCO as well as TOMIE data on the weights of the best epoch. For LOCO and TOMIE the results are listed in Table 2.

Table 2: Inference results one the LOCO base model with LOCO and TOMIE.

Class Dataset	All	Small load carrier	Forklift	Pallet	Stillage	Pallet truck
LOCO	0.52	0.467	0.350	0.737	0.688	0.359
TOMIE	0.234	0.0	0.0	0.023	0.422	-

For fine-tuning with TOMIE, subset B was selected as training set and subset D as validation set. After conducting runs with several optimizer and freeze parameters (T), the best mAP@.5 reached was 0.67 using SGD and no frozen layers (T0). To prevent catastrophic forgetting, we applied mixed fine-tuning (M) and frozen layers: LOCO subset 1 was added to the training set, while subset 4 was added to the validation set, since both were not used to train the base model. The optimizer used was SGD and the LR was increased from 0.001 to 0.01 due to a higher number of and more diverse training data. Results are shown in Table 3.

Table 3: Inference results (mAP@.5) for TOMIE and LOCO with LOCO based model fine-tuned with TOMIE.

Inference results for TOMIE subsets A, C, E and F on the LOCO base model fine-tuned with TOMIE					Inference results for LOCO subset 4 on the LOCO base model fine-tuned with TOMIE			
Class	M T50	M T5	T0	T50	M T50	M T5	T0	T50
All	0.65	0.645	0.630	0.624	0.182	0.006	0.033	0.001
Forklift	0.35	0.289	0.293	0.338	0.057	0.001	0.002	0.001
Pallet	0.608	0.65	0.656	0.562	0.552	0.022	0.127	0.002
Stillage	0.993	0.995	0.995	0.972	0.036	0	0	0
Pallet truck	-	-	-	-	0.081	0	0.001	0

5.1.2 TOMIE Base Model

SGD was chosen, with an LR of 0.01 and five frozen layers as well as a resolution of 1152px. Subsets A, B, E and F served as the training set, while subset C and D were used for validation. After the 10th epoch a training mAP@.5 of 0.846 was reached, which was not surpassed after. Inference was first conducted with the LOCO benchmark (subsets 1 and 4) on the weights from the best epoch, which lead to an mAP@.5 of 0.002. On the TOMIE base model, fine-tuning with LOCO was conducted for 40 epochs while using SGD: once with 10, the other time with no frozen layers. The results are shown in Table 4.

Table 4: Inference results mAP@.5 for LOCO subsets 1 and 4 on the TOMIE base model.

Class Dataset	All	Small load carrier	Forklift	Pallet	Stillage
LOCO	0.002	0.0	0.0	0.001	0.005
TOMIE	0.839	0.826	0.613	0.937	0.977

The split and the other parameters are the same as before. The mAP@.5 inference results for both trained models are shown in Table 5, each for the weights of the best epoch.

Table 5: Inference results mAP@.5 on the TOMIE base model fine-tuned with LOCO.

Class	LOCO (T10)	TOMIE (T10)	LOCO(T0)	TOMIE(T0)
All	0.491	0.243	0.514	0.215
Small load carrier	0.379	0	0.485	0
Forklift	0.258	0	0.266	0
Pallet	0.737	0.011	0.745	0.019
Stillage	0.652	0.961	0.663	0.841
Pallet truck	0.431	-	0.411	-

5.2 Synthetic Data

In this section we analysed the quality of our SYNTH dataset by training models only with the pallet class. For this, we trained a SYNTH base model for 40 epochs with five frozen layers and SGD with a LR of 0.01. Subsets α and γ were used as training set, subset ϵ as validation set, the remaining data were used for inference testing. This way, the subsets with the most DR are used for training and the one closest to reality for validation. Proceeding with our SYNTH base model, we conducted additional experiments by fine-tuning the model with TOMIE and LOCO. We also trained a two staged fine-tuned SYNTH base model, which was first fine-tuned with TOMIE and LOCO afterwards. For inference testing we used subset 1 from LOCO. All experiment results are shown in Table 6.

Table 6: Inference results (mAP@.5) with LOCO subset 1 on different configurations.

Model	mAP@.5
SYNTH	0.096
LOCO fine-tuning on SYNTH	0.740
TOMIE fine-tuning on SYNTH	0.005
LOCO fine-tuning on TOMIE on SYNTH	0.724

In addition to our fine-tuned experiments we also tested all subsets from LOCO and TOMIE to determine how the inference with real data is when used only with models trained using SYNTH, which is shown in Table 7.

Table 7: Inference results on the SYNTH base model.

LOCO subset	1	2	3	4	5		all
mAP@.5	0.113	0.044	0.029	0.054	0.054		0.051
TOMIE subset	A	B	C	D	E	F	all
mAP@.5	0.718	0.263	0.416	0.003	0.752	0.015	0.343

6. Evaluation

6.1.1 Real Data

The LOCO base model shows an improved mAP compared to the original paper. However, while pallets and stillages produce reliable results, small load carriers, forklifts and pallet trucks perform poorly. For forklifts and pallet trucks, the low number of occurrences and variety of the objects are likely to be responsible for the model's poor performance. Delving deeper into LOCO, one reason is the annotation quality (see Figure 3). There are missing and faulty annotations, when comparing inference with ground truth. Also, the objects are in part difficult to recognise, be it due to their size or blur.

Figure 3: Comparison of the ground truth (left) and the predictions (right) in LOCO.

For the TOMIE base model, we discovered a fast training convergence. One possible reason for this is that the data is quite easy to learn on. The stillage class is detected well, which is because in all images it is the very same stillage instance. The performance for the forklift is, in comparison, rather poor. A reason for this might be the inconsistent annotations (see Figure 4) with sometimes the person dragging it being included and sometimes not. Annotation inconsistency also occur for other categories, which could lead to the notably fluctuation in training performance. Also, the model is overfitting on loaded pallets, which are sometimes detected even if they are occluded.

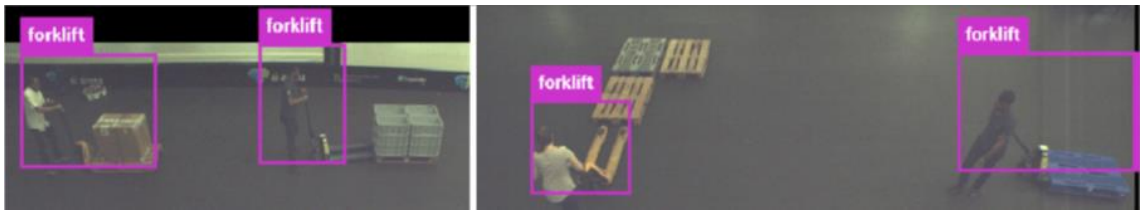

Figure 4: Examples of annotation inconsistencies of forklifts in TOMIE.

The performance of TOMIE data on the LOCO model is poor, except for the stillage class. Vice versa, performance is even worse. A central criteria is probably the camera perspective. In TOMIE it is almost a top-down, bird's eye view, whereas in LOCO the pictures were all taken relatively close to the ground. In addition, a model trained on LOCO mostly labels the visible part of loaded pallets. With TOMIE, the bounding box always covers the whole pallet. This leads to incorrect predictions due to low IoU, even though the pallet may have been recognised correctly (see Figure 5).

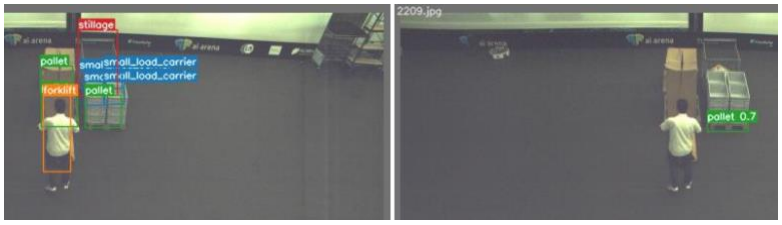

Figure 5: Pallet annotation compared to inference results with LOCO on the TOMIE dataset.

6.1.2 Synthetic Data

The SYNTH model performs much better on TOMIE than on LOCO, showing that the reality gap could be overcome with synthetic data. A deeper investigation of the LOCO dataset shows that pallets in shelves are rarely recognised, especially if the pallets do not face the camera or the instance is small (Figure 6, bottom left). The LOCO base model can recognise these objects much better. Stacked pallets are also problematic (Figure 6, top images).

Figure 6: Inference samples with LOCO subset 1 (dark green = ground truth, orange = LOCO base model, blue = SYNTH base model).

For pallets, in general it seems like fine-tuning with LOCO on a base model improves performance and again fine-tuning on SYNTH performs slightly better than fine-tuning on TOMIE. For fine-tuning with TOMIE on SYNTH, the inference with subset 1 performs very poorly as usual and as expected, fine-tuning on LOCO still brings a better performance. Finally, we analysed the combination of the SYNTH base model with LOCO and TOMIE data. For inference with subset 1 with LOCO fine-tuned model, a better performance can be achieved by fine-tuning.

7. Conclusion

In this paper, we conducted a thorough examination of two warehousing logistics datasets, LOCO and TOMIE. Our investigation not only involved a comparative analysis but also sought broader insights relevant to computer vision and logistics object detection. While first results demonstrated the superiority of our models on the LOCO benchmark, we identified inherent dataset limitations, including class distribution imbalances and annotation inaccuracies, particularly with pallets and small load carriers. TOMIE, with its consistent camera perspectives, presented challenges like label inconsistencies and fluctuations during training. These issues underscore the importance of robust dataset curation and annotation. Secondly, our analysis revealed substantial disparities between LOCO and TOMIE, spanning different camera perspectives, environmental conditions, and labelling approaches, shown by catastrophic forgetting during transfer learning. To bridge these gaps, we introduced a synthetic data generation pipeline, effectively leading to performance enhancements. In summary, our study not only offers insights into LOCO and

TOMIE but also underscores the broader relevance of robust dataset creation and domain adaptation challenges in logistics object detection. Bridging these gaps is crucial for enhancing model robustness and applicability in real-world logistics scenarios.

References

[1] Li, T., Huang, B., Li, C., Huang, M., 2019. Application of convolution neural network object detection algorithm in logistics warehouse. The Journal of Engineering 2019.

[2] Mayershofer, C., Ge, T., Fottner, J., 2021. Towards Fully-Synthetic Training for Industrial Applications, in: , LISS 2020, pp. 765–782.

[3] Mayershofer, C., Holm, D.-M., Molter, B., Fottner, J., 2020. LOCO: Logistics Objects in Context, in: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 612–617.

[4] Jérôme Rutinowski, Hazem Youssef, Sven Franke, Irfan Fachrudin Priyanta, Frederik Polachowski, Moritz Roidl, Christopher Reining, 2023. Semi-Automated Computer Vision based Tracking of Multiple Industrial Entities - A Framework and Dataset Creation Approach.

[5] Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.M., 2022. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. https://arxiv.org/pdf/2207.02696.

[6] Hafiz Mughees Ahmad, Afshin Rahimi, 2022. Deep learning methods for object detection in smart manufacturing: A survey. Journal of Manufacturing Systems 64, 181–196.

[7] Zaccaria, M., Monica, R., Aleotti, J., 2020. A Comparison of Deep Learning Models for Pallet Detection in Industrial Warehouses, in: 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 417–422.

[8] Manurung, P., Rusmin, P.H., Yusuf, R., 2022. Custom Pallet Detection Using YOLOv5 Deep Learning Architecture, in: 2022 International Symposium on Electronics and Smart Devices (ISESD), pp. 1–6.

[9] Li, Y., Ding, G., Li, C., Wang, S., Zhao, Q., Song, Q., 2023. A systematic strategy of pallet identification and picking based on deep learning techniques. Industrial Robot: the international journal of robotics research and application (ahead-of-print).

[10] Stefan Hinterstoisser, Olivier Pauly, Hauke Heibel, Martina Marek, Martin Bokeloh, 2019. An Annotation Saved is an Annotation Earned: Using Fully Synthetic Training for Object Instance Detection.

[11] Markus Knitt, Jakob Schyga, Asan Adamanov, Johannes Hinckeldeyn, Jochen Kreutzfeldt, 2022. Estimating the Pose of a Euro Pallet with an RGB Camera based on Synthetic Training Data.

[12] Sergey I. Nikolenko, 2019. Synthetic Data for Deep Learning.

[13] Rutinowski, J., Vankayalapati, B., Schwenzfeier, N., Acosta, M., Reining, C., 2022. On the Applicability of Synthetic Data for Re-Identification. https://arxiv.org/pdf/2212.10105.

[14] Akar, C.A., Tekli, J., Jess, D., Khoury, M., Kamradt, M., Guthe, M., 2022. Synthetic Object Recognition Dataset for Industries, in: 2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 150–155.

Biography

Dimitrij-Marian Holm (*1990) studied Robotics at the University of Applied Science Munich. Since 2018 he is part of the Research Staff at the Chair of Materials Handling, Material Flow, Logistics, focusing on Machine Vision and Learning.

Philipp Junge (*1997) received a B.Sc. in Mechanical Engineering before completing the M.Sc. in Mechatronics & Robotics at the TU Munich in 2023. In his Master's thesis he worked at the Chair of Materials Handling, Material Flow, Logistics on the topic presented in this publication.

Jérôme Rutinowski (*1996) studied Mechanical Engineering at Ruhr-University Bochum. Since 2020 he is a researcher and PhD candidate at Prof. Dr. Dr. h. c. Michael ten Hompel's Chair of Material Handling and Warehousing at TU Dortmund University.

Johannes Fottner received a Dr.-Ing. degree in mechanical engineering from the Technical University of Munich (TUM), Munich, Germany in 2002. During the past 15 years, Johannes Fottner acted in different managing functions in the material-handling sector. Since 2016, he has been the head of the Chair of Materials Handling, Material Flow, Logistics with the TUM.