Pose-aware object recognition on a mobile platform via learned geometric representations*

Csaba Beleznai, Philipp Ausserlechner, Andreas Kriegler, Wolfgang Pointner
Center for Vision, Automation & Control
AIT Austrian Institute of Technology
Vienna, Austria
csaba.beleznai@ait.ac.at

***Abstract*—Mobile robot operations are becoming increasingly sophisticated in terms of robust environment perception and levels of automation. However, exploiting the great representational power of data-hungry learned representations is not straightforward, as robotic tasks typically target diverse scenarios with different sets of objects. Learning specific attributes of frequently occurring object categories such as pedestrians and vehicles, is feasible since labeled data-sets are plenty. On the other hand, less common object categories call for the need of use-case-specific data acquisition and labelling campaigns, resulting in efforts which are not sustainable with a growing number of scenarios. In this paper we propose a structure-aware learning scheme, which represents geometric cues of specific functional objects (airport loading ramp) in a highly invariant manner, permitting learning solely from synthetic data, and also leading to a great degree of generalization in real scenarios. In our experiments we employ monocular depth estimation for generating depth and surface normal data and in order to express geometric traits instead of appearance. Using the surface normals, we explore two different representations to learn structural elements of the ramp object and decode its 3D pose: as a set of key-points and as a set of 3D bounding boxes. Results are demonstrated and validated in a series of robotic transportation tasks, where the different representations are compared in terms of recognition and metric space accuracy. Te proposed learning scheme can be also easily applied to recognize arbitrary man-made functional objects (e.g. containers, tools) with and without known dimensions.**

***Index Terms*—robot vision, environment perception, geometric cue learning, monocular depth**

I. Introduction

Autonomous robot operations are emerging in various unconstrained environments such as in logistics, construction and agriculture. Environment perception and spatial reasoning are important technology components enabling these applications. In recent years Deep Learning has substantially advanced the state of several core technologies, where the object detection and pose estimation tasks reach a high accuracy in diverse settings. In these tasks learned representations are used to establish a perceived environment model in terms of a set of pre-defined entities and their spatial relations.

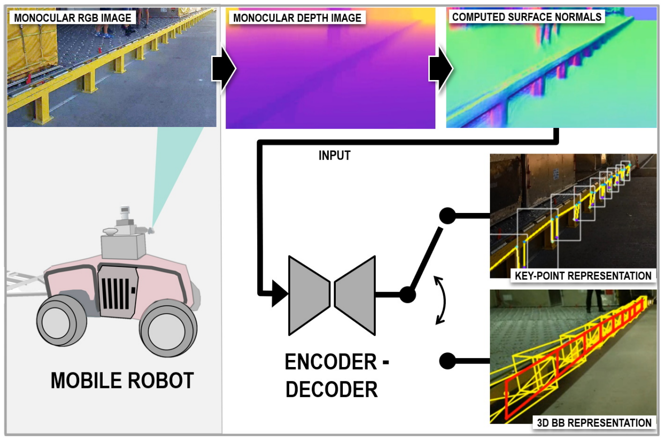

Fig. 1. Illustration depicting the employed input image representation and the two distinct inference tasks for the use-case of approaching a loading ramp.

Despite the significant progress, learning robust representations is often hindered by the need of exhaustive learning of vast set of object appearances under different photometric conditions and varying view geometries. This requirement implies the availability of large and diverse annotated datasets. For less common object categories, however, learning schemes often must rely on curated data collection or synthetic data simulation. Certain input representations, such as geometric cues, also offer ways to accomplish more data-efficient learning. Most importantly, instead of learning all possible object appearances, learning geometric representations discards photometric and appearance variations and it captures a simpler innate object property. This reduced representational space implies less data needed for learning. The emergence of enhanced depth sensing modalities such as high-quality stereo vision, monocular depth estimation, LiDAR, Radar also supports this research direction and can lead to data representations highly invariant with respect to view, appearance and photometric variations.

In this paper we aim at an environment perception functionality for a mobile robot, enabling it to approach and align itself to a pre-defined elongated structure (loading ramp) for assisting ground cargo handling. To this end, we present a generic structure-aware learning scheme to estimate the 3D pose of man-made functional objects (via the example of the airport loading ramp) from a monocular view, by exploiting recent advances of monocular depth estimation [1]

*This work was carried out in the AUTILITY project, supported by the "ICT of the Future" programme of the Austrian Research Promotion Agency (FFG) under the project nr. 867556.

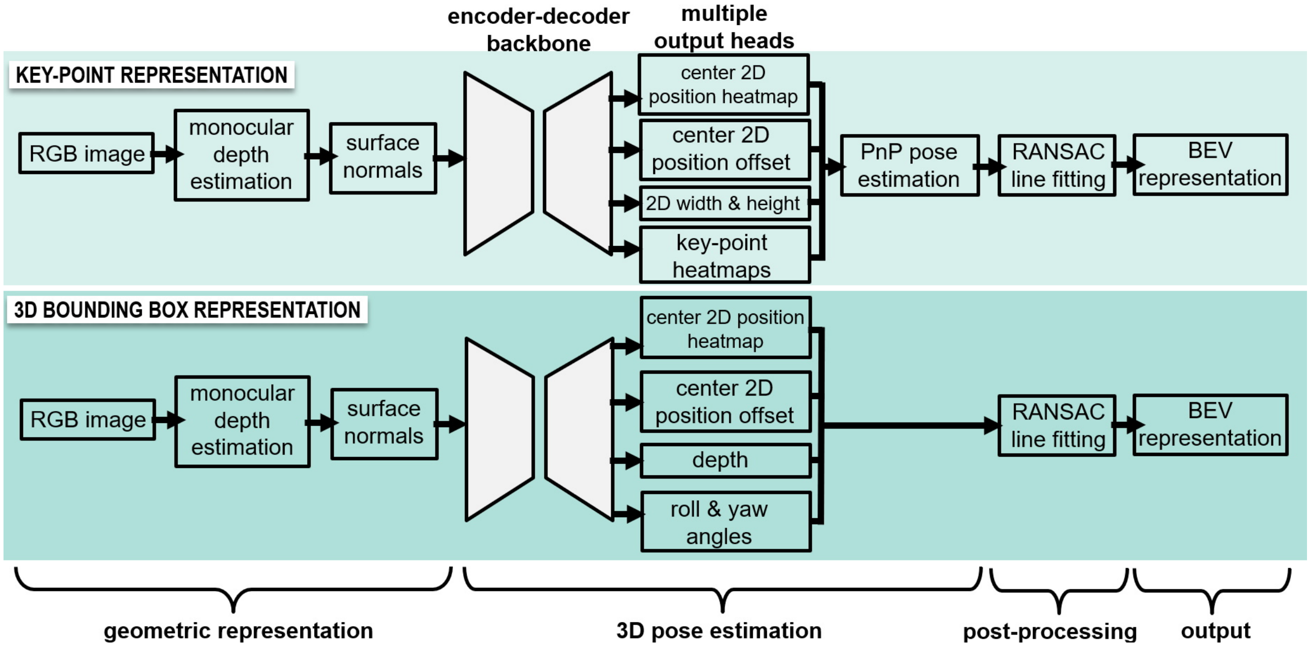

Fig. 2. Overview of the two representational (key-point and 3D-BB) and pose estimation pipelines. The neural network input in both cases is an image of the surface normals, computed from a monocular depth estimate.

and learning. We employ the computed surface normals of the depth image (see Fig. 1), which introduces two advantages: First, the surface normal image encodes scale-invariant local geometric cues, which allows representing local structural elements, such as 3D edges, corners, local structural patterns. Furthermore, when generating synthetic data of man-made structures, monocular depth and surface normal estimation yields results where the data quality gap between synthetic and real is small. These two representational traits allow us to formulate a pose estimation learning task purely learned from synthetic data and well generalizing in real scenarios. We explore two independent learning pipelines, key-point- and 3D-bounding-box-based object representations to derive object detection hypotheses in a metric birds-eye-view (BEV) space (see fig. 2). The motivation behind exploring and comparing these two representations is given by their different ways to represent spatial relations and correlations. Key-points are of more localized spatial extent, therefore a robustness in case of partial object visibility is expected. On the other hand, 3D bounding box (3D-BB) regression learns an object-holistic representation, presumably contributing to a more accurate spatial localization. The proposed scheme is simple, and it can be applied to recognize arbitrary man-made objects with distinct local structural elements. Monocular depth estimation can also be replaced by stereo depth computation.

The remainder of the paper is structured as follows: section II gives an overview on related work. Section III describes the data generation and pose estimation methodology for both key-point- and 3D-bounding-box-based schemes. Section IV presents experimental results and a discussion in light of a real robotic setup. Finally, Section V concludes the paper.

II. Related work

3D object pose denotes the spatial transform needed to align the coordinate reference of an observed object with that of the observer. Therefore, accurate object detection and pose estimation are key vision tasks enabling robotic spatial reasoning and manipulation. The Amazon Picking Challenge [2] is a relevant example where recovering object pose from a close object set allows for automated part manipulation. Industrial part recognition [3] and household robotic manipulation [4] are also relevant applied examples. Recent reviews on 3D pose estimation [5], [6] provide a comprehensive overview both on the representational and applied aspects.

Recent research activities focus on learned representations of geometric nature. This emerging field of geometric deep learning is well summarized in [7], [8], where geometric principles are highlighted to explain regularities often observed in the physical world. Depth data naturally conveys geometric information, therefore understanding depth computation, its data characteristics and its failure modes are highly pertinent. [9] outlined four steps commonly encountered in classical stereo image pipelines. Despite representational advances via Deep Learning, these steps continue to play a key role [10]. Depth estimation from a single image, also denoted as monocular depth estimation, has recently emerged as an appealing alternative to depth estimation from stereo image pairs [11]. An enhanced generalization of monocular depth models is attained via a mixture of datasets in [12]. Recent representational advances based on Vision Transformers [13], exploiting the attention-mechanism [14] are capable to accurately capture long-range semantic relations [15], see also [16] for a survey.

The 3D pose estimation task spans a 6D search space of

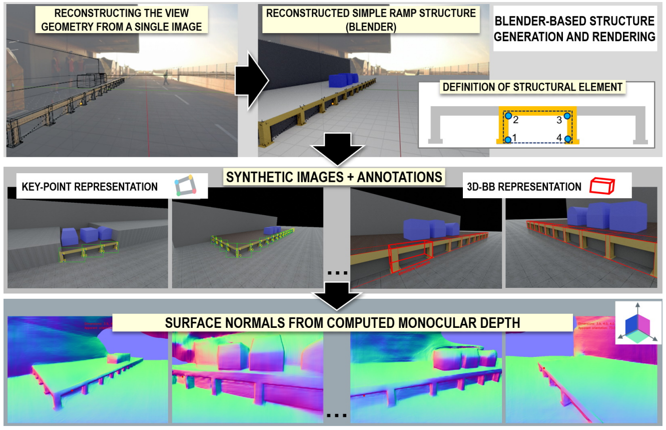
RECONSTRUCTING THE VIEW GEOMETRY FROM A SINGLE IMAGE
RECONSTRUCTED SIMPLE RAMP STRUCTURE (BLENDER)
BLENDER-BASED STRUCTURE GENERATION AND RENDERING
DEFINITION OF STRUCTURAL ELEMENT
2
3
1
4
KEY-POINT REPRESENTATION
SYNTHETIC IMAGES + ANNOTATIONS
3D-BB REPRESENTATION
...
SURFACE NORMALS FROM COMPUTED MONOCULAR DEPTH
...

Fig. 3. The proposed training data generation pipeline: first, a simple initial view and scene geometry is created in Blender (top). Random view and structural variations are used to generate synthetic color data with two different annotations (key-points, 3D-BB). The rendered color images are used for monocular depth estimation and surface normal computation (bottom).

object location $\mathbf{x} = (x, y, z)$ and orientation $\theta = (\theta_r, \theta_p, \theta_y)$. Many methods rely on RGB-D images to recover these parameters. Representations have progressed from handcrafted schemes, e.g. Linemod [17], to end-to-end learned pose estimation schemes [18], [19], [20]. Most common approaches formulate the learning task [18], [21], [19] as corner point regression, followed by a PnP step [22]. The proposed direct regression scheme of 3D bounding box (3D-BB) pose parameters in [23] has triggered a large body of works [24], [25] focusing on 3D-BB pose regression from a monocular image or from depth data.

Our proposed methodology shares many representational aspects (pose parametrization, pose computation via point regression) with existing works, however, it exploits the intermediary monocular depth estimation step in a novel way to derive a geometry-aware learning scheme generalizing from synthetic training data.

III. Methodology

In our proposed learning scheme, as shown in fig. 2, we explore two independent object representations both yielding a 3D pose estimate. Both representational pipelines rely on an input map of surface normals, computed from a monocular depth estimate. In the followings we describe the synthetic data generation and learning steps in detail.

A. Data generation

Synthetic data generation allows for creating vast quantities of geometric structural variations which, when rendered, result in images depicting structures from diverse view-points. Using the object appearance directly from the color images for learning is hindered by the fact that models trained on synthetic data often exhibit a severe accuracy degradation when facing the real data domain. Generating photo-realistic models is a remedy, but it requires much effort as appearance often exhibits a great diversity. Depth data, on the other hand, significantly reduces variations and geometric cues become the prevailing information. Monocular depth estimation [15] correctly infers local geometric structures (e.g. 3D edges and corners), which nevertheless become less and less accurate at a larger spatial scale (e.g. warped ground plane). Monocular depth estimation also introduces smoothing and noise artifacts, which prove to be beneficial, as depth estimates from real images exhibit the same discrepancies, thus lowering the gap between simulated and real data.

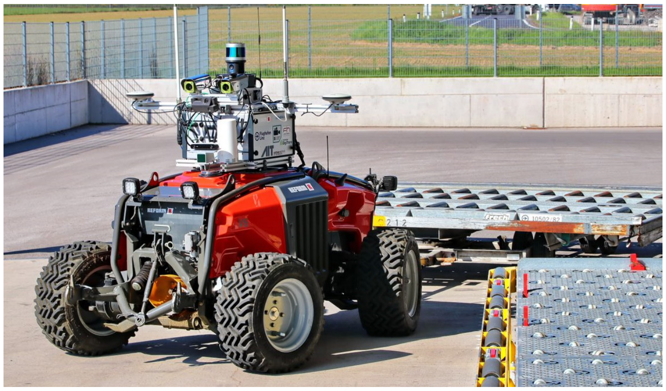

Fig. 4. The employed mobile platform (Reform-Werke Metron P48 RC) equipped with multi-sensor and control hardware.

As shown in fig. 3 we first re-create the vehicle on-board camera's view geometry. To this end, we use a single on-board image of the scene, Blender [26] and the publicly available fSpy toolkit [27]. This tool employs a photogrammetric scheme [28] to estimate the camera view geometry and its focal length from a single photo depicting our ramp structure with known dimensions. After establishing a view geometry, we manually create an initial simple scene with an arbitrary number of ramp structural elements and also distracting ramp-like objects (3D blocks) to force representation learning to focus on ramp-specific structural elements (e.g. T-junctions). Programmatically, using Blender's python API, we randomize the camera viewpoint within a range typical for our use-case. Furthermore, we compute for each ramp structural unit annotations which unambiguously describe their structure via a set of key-points and 3D bounding box pose parameters. The rendered color images contain sufficient texture and shading information encoding the perspective such that a subsequent monocular depth estimation step [15] can estimate a consistent depth representation (not shown). Next, we use a simple procedure to transform depth images to surface normals. We calculate depth derivatives along x and y using the Sobel kernel. Using the computed gradients we build local support planes, whose normal vectors can be seen as the normal vectors of the object surface at those pixels. The directional vector components of the computed surface normals are mapped to respective 8-bit RGB channels (see fig 3). In this way we generate 100K images with a resolution of 768×512 px and corresponding annotations for training.

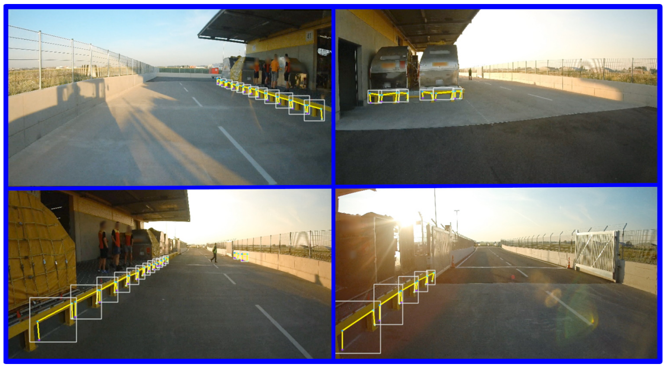

Fig. 5. Example results for detecting ramp structural elements (as defined in Fig. 3) via a set of key-points. Note the varying illumination and viewing conditions.

B. Pose-aware detection via key-points

The computed surface normals and corresponding annotations represent the input of our learning scheme. Both key-point and 3D-BB representation learning methods employ the CenterNet [25] architecture as the basis. We use the DLA-34 hierarchical layer fusion network scheme to enlarge the spatial scope considered during learning. We consider each ramp structural element (see fig. 3) as an object. In an image containing n object instances, we seek to estimate following two-dimensional attributes: $\{(\hat{x}_i, \hat{y}_i), (\delta\hat{x}_i, \delta\hat{y}_i), (\hat{w}_i, \hat{h}_i)\}_{i=1}^{n}$, corresponding to the integer-valued object center, float-valued 2D offsets and 2D bounding box dimensions, respectively. We define 4 key-points representing an object and similarly to human key-point estimation in [25] we estimate $4n$ key-point locations and key-point-to-center offsets. Given the known dimensions of sought structure 3 and the 4 estimated key-points defining a 3D planar structure, a PnP scheme can be applied to lift the 2D estimates into 3D.

C. Pose-aware detection via 3D-BB regression

3D pose estimation requires the estimation of 3D translational and rotational parameters. We formulate the set of sought parameters as a hybrid, 2D and 3D regression task, while employing the same representational backbone as in the previous case. Each ramp structure unit (denoted as object) has a 3D center estimate: $\left[\hat{X}\ \hat{Y}\ \hat{Z}\right]^{\intercal}$. The regression task of this variable becomes substantially easier, if the directly observable 2D center position $(\hat{x}, \hat{y})$ and the object distance (depth) $\hat{Z}$ are formulated as estimation tasks. Upon an estimated 2D center position and corresponding depth, the 3D translational parameters can be obtained by:

$$\hat{X} = (\hat{x} - p_x) * \hat{Z}/f_x\ , \quad \hat{Y} = (\hat{y} - p_y) * \hat{Z}/f_y \qquad (1)$$

where f_x, f_y, p_x, p_y are the focal lengths and principal points along the x and y image plane axes, respectively. We consider two rotational roll and yaw angles (α, β) for each structural unit. During data generation the observed angles (α_v, β_v) with respect to the camera are computed via:

$$\alpha_v = \alpha + \arctan(X/Z)\ , \quad \beta_v = \beta + \arctan(Y/Z) \qquad (2)$$

in order to establish a consistent correlation between the observable orientations with the data patterns in the normal images. The two orientation are encoded as $[\sin(\alpha_v), \cos(\alpha_v)]^{\intercal}$ and $[\sin(\beta_v), \cos(\beta_v)]^{\intercal}$ for learning. As the ramp structural units exhibit mirror symmetry, we only consider rotations within the $[0, \pi]$ range.

For both representational cases, we formulate a composite loss function where key-point classification employs a focal loss term [25], while the other parameters are penalized via an L1 loss function. In all cases we consider a single object class.

D. Post-processing: 3D structure fitting

For key-point-based object proposals, the proposals must be lifted into the 3D space. As the 4 key-points define a planar structure of known dimensions, we employ the ePnP [29] scheme to compute corresponding 3D coordinates of a 3D plane, thus placing the structure into a birds-eye-view (BEV) space.

In order to enforce the structural constraint of a linear structure, we employ a RANSAC-based line-fitting scheme operating on the set of proposed structure elements in the BEV space. We treat the individual proposed elements as oriented line segments and we compute a fitness score of tentative matches based as 3D spatial proximity and angular alignment. Since the *inlier-vs-outlier* ratio in all 25 approach maneuvers was very high, therefore the best matching ramp fit could be found in an unambiguous and temporally stable manner. No structure tracking was applied, but it would represent a straightforward extension.

IV. Results and Discussion

The mobile robotic setup consisted of a Metron P48RC [30] vehicle equipped with an RTK GPS, a LeiShen surround LiDAR and a stereo camera of 3.2 Mpixel resolution. The proposed methodology used only one view of the stereo setup, after resizing the image to 768×512 pixels. The on-board unit was an NVidia Xavier AGX platform running directly the ramp pose detection code in PyTorch. Both detection and pose estimation pipelines run at about 7 fps without any run-time optimization. The vehicle was used in 3 dataset recording campaigns under different seasonal and time-of-the-day conditions. 25 test datasets (consisting of more than 75000 frames) have been selected for evaluation, where ramp drive-by's and approaches along different paths of travel have been carried out.

Results for key-point-based detection are shown in fig. 5. The true ramp structural elements have been detected with a high recall under varying viewing conditions, even at rather small scales (down to 16 px). However, certain step-like structures (fence, barrier) have occasionally caused isolated false detection responses, despite representing such clutter objects during data generation.

Results for 3D-BB-based detection are shown in fig. 6. The top row of the figure also shows the computed surface normals for two different illumination conditions. As it can

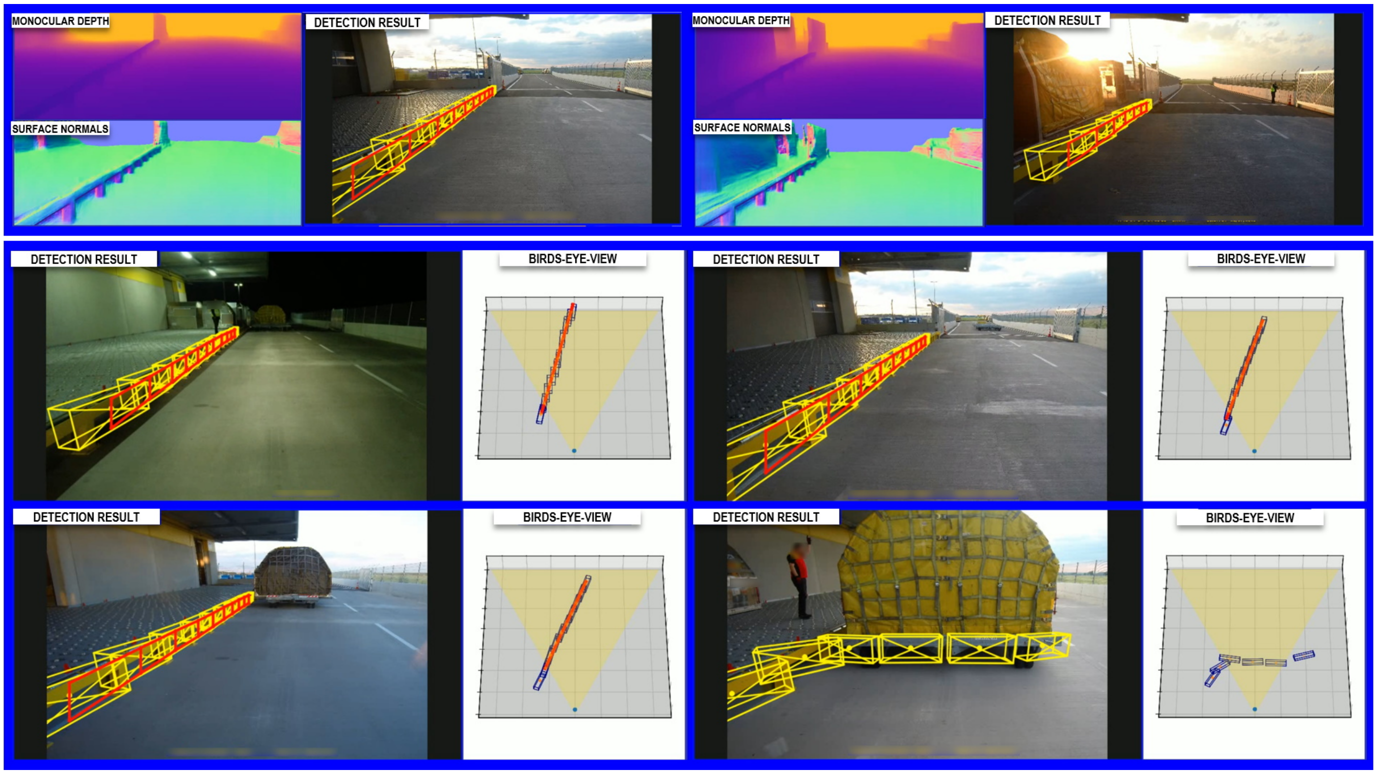

Fig. 6. Overview on the employed input data and accomplished results for ramp detection via 3D-BB regression. Top row: Monocular depth image, its computed normals and detection results for two different illumination conditions. Center and bottom rows: detection results for four different scenarios displayed in the image and birds-eye-view space. The red line implies the detected ramp structure. The last image shows the achieved generalization towards generic ramp-like objects.

be seen, the structure of surface normals accurately reflects the local ramp geometry, well outlining edges and junctions despite low-contrast or different illumination conditions. 3D-BB results indicate that the pose of the ramp structural units can be more accurately recovered than via key-points. 3D bounding boxes enforce more object-like representation, as the learning task targets an oriented bounding box estimate best matching the data. Key-point-based representation is more local in its nature, as its structure is defined by edges connecting local point estimates. This characteristics results in more structural flexibility and also inaccuracy. The higher spatial accuracy of the 3D-BB representation can be observed in all results showing back-projected structures. Furthermore, the last image depicts the high generalization ability of the proposed representation towards similar ramp structures, as the ramp structure units of a carrier vehicle (same height, similar steel frame) also accurately detected.

TABLE I
EVALUATION OF DETECTION AND POSE ESTIMATION ACCURACY

Measure	REPRESENTATION	
	key-point	*3D bounding box*
RMSE	0.97	0.73
Precision	0.51	0.64
Recall	0.48	0.52

We perform a simple quantitative evaluation to examine the spatial accuracy of the pose estimates, and the recognition accuracy. To generate metric ground truth, we select 25 image frames (one representative frame of each test run) with a ramp structure. Using the same photogrammetry pipeline via fSpy, as used for the data generation step of section III-A, we manually place a set of structural elements in the 3D space (aligned to the image structures) to create metric ground truth in the camera space. Using the root-mean-square-error (RMSE) (eq. 3) between N ground truth (X_i) and estimated ($\hat{X}_i$) ramp unit 3D centers, we compute a measure for the spatial accuracy of estimates. The accuracy values for both key-point and 3D-BB representations are shown in table I.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{X}_i)^2} \tag{3}$$

To quantify the recognition accuracy of our 1-class problem (ramp structural unit vs. background), we employ a common evaluation procedure used in works targeting 3D detection and localization [24]. Using the manually created 3D bounding box ground truth for each structural element, we compute the 3D IoU with a threshold of 0.7. Based on this overlap-based association measure, we compute precision and recall values for both representations, as shown in table I. As it can be seen from the table, 3D-BB-representation yields a higher spatial accuracy in terms of 3D RMSE error. In terms of recognition accuracy, this representation also seems to be advantageous, as it exhibits (upon a detected structural element) a higher spatial accuracy, although suffers from a sensitivity to occlusion and failure for small-sized objects. From an applied point-of-view, both representations meet the task requirement, as the ramp structure is detected in all 25 test sequences, allowing for subsequent spatial reasoning.

V. Conclusions

In this paper we present a simple learning scheme applicable to robotic scenarios where object recognition of less common objects without the need of manual annotations is needed. The proposed scheme demonstrates that learned geometric representations can unambiguously localize and recover the 3D pose of these objects, and despite the training from the purely synthetic domain, the model generalizes extremely well in the real domain. The paper also presents experimental evidence that depth data from monocular (and also stereo) depth estimation conveys sufficiently accurate spatial information to perform occlusion- and illumination-robust pose estimation. Future work will investigate the use and advantages of stereo depth over monocular depth estimates.

VI. Acknowledgements

We gratefully acknowledge the support by Marcel Zeilinger (AIT) in the data acquisition and on-board testing tasks. We are also indebted to Reform-Werke GmbH for their support and development of the robotic hardware.

References

[1] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *arXiv preprint arXiv:1907.01341*, 2019.

[2] Y. Domae, "Amazon Picking Challenge 2016," *Journal of the Robotics Society of Japan*, 2016.

[3] C. Sahin and T. K. Kim, "Recovering 6D object pose: A review and multi-modal analysis," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019.

[4] R. B. Rusu, Z. C. Marton, N. Blodow, M. Dolha, and M. Beetz, "Towards 3D Point cloud based object maps for household environments," *Robotics and Autonomous Systems*, 2008.

[5] C. Sahin, G. Garcia-Hernando, J. Sock, and T. K. Kim, "A review on object pose recovery: From 3D bounding box detectors to full 6D pose estimators," 2020.

[6] S. Hoque, M. Y. Arafat, S. Xu, A. Maiti, and Y. Wei, "A Comprehensive Review on 3D Object Detection and 6D Pose Estimation with Deep Learning," *IEEE Access*, 2021.

[7] M. Bronstein, J. Bruna, T. Cohen, and P. Velivckovic, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," *ArXiv*, vol. abs/2104.13478, 2021.

[8] W. Cao, Z. Yan, Z. He, and Z. He, "A Comprehensive Survey on Geometric Deep Learning," *IEEE Access*, vol. 8, pp. 35 929–35 949, 2020.

[9] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," Microsoft Research, Tech. Rep., 2001.

[10] K. Zhou, X. Meng, and B. Cheng, "Review of Stereo Matching Algorithms Based on Deep Learning," *Computational Intelligence and Neuroscience*, vol. 2020, no. 1, 2020.

[11] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F. Qian, "Monocular depth estimation based on deep learning: An overview," *Science China Technological Sciences*, pp. 1–16, 2020.

[12] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 14, 2020.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv:2010.11929*, p. 21, 2020.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017, pp. 6000–6010.

[15] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," *arXiv:2103.13413*, p. 15, 2021.

[16] S. Khan, M. Naseer, M. Hayat, S. Waqas Zamir, F. Shahbaz Khan, and M. Shah, "Transformers in vision: A survey," *ArXiv:2101.01169*, p. 28, 2021.

[17] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013.

[18] B. Tekin, S. N. Sinha, and P. Fua, "Real-Time Seamless Single Shot 6D Object Pose Prediction," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[19] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4561–4570.

[20] K. Park, T. Patten, and M. Vincze, "Pix2pose: Pixel-wise coordinate regression of objects for 6D pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[21] G. Pitteri, S. Ilic, and V. Lepetit, "Cornet: generic 3d corners for 6d pose estimation of new objects without retraining," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[22] Y. Wu and Z. Hu, "Pnp problem revisited," *Journal of Mathematical Imaging and Vision*, vol. 24, no. 1, pp. 131–141, 2006.

[23] A. Mousavian, D. Anguelov, J. Košecká, and J. Flynn, "3D bounding box estimation using deep learning and geometry," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

[24] J. Ku, A. D. Pon, and S. L. Waslander, "Monocular 3D object detection leveraging accurate proposals and shape reconstruction," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019.

[25] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as Points," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[26] *Blender - a 3D modelling and rendering package*, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. [Online]. Available: http://www.blender.org

[27] Stuffmatic, "fSpy." [Online]. Available: https://fspy.io

[28] E. Guillou, D. Méneveaux, E. Maisel, and K. Bouatouch, "Using vanishing points for camera calibration and coarse 3d reconstruction from a single image," *Vis. Comput.*, vol. 16, no. 7, pp. 396–410, 2000.

[29] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate O(n) solution to the PnP problem," *International Journal of Computer Vision*, 2009.

[30] Reform-Werke. (2021) Metron p48 rc. Reform-Werke Bauer and Co Gesellschaft m.b.H. Labs. [Online]. Available: https://www.reform.at/produkte/metron/metron-p48