Evaluation of Monocular and Stereo Depth Data for Geometry-Assisted Learning of 3D Pose

Andreas Kriegler [1,2], Csaba Beleznai [1] and Margrit Gelautz [2]

***Abstract*— The estimation of depth cues from a single image has recently emerged as an appealing alternative to depth estimation from stereo image pairs. The easy availability of these dense depth cues naturally triggers research questions, how depth images can be used to infer geometric object and view attributes. Furthermore, the question arises how the quality of the estimated depth data compares between different sensing modalities, especially given the fact that monocular methods rely on a learned correlation between local appearance and depth, without the notion of a metric scale. Further motivated by the ease of synthetic data generation, we propose depth computation on synthetic images as a training step for 3D pose estimation of rigid objects, applying models on real images and thus also demonstrating a reduced synth-to-real gap. To characterize depth data qualities, we present a comparative evaluation involving two monocular and one stereo depth estimation schemes. We furthermore propose a novel and simple two-step depth-ground-truth generation workflow for a quantitative comparison. The presented data generation, evaluation and exemplary pose estimation pipeline are generic and applicable to more complex geometries.**

I. INTRODUCTION

Recent scientific trends increasingly allow for an enhanced spatial perception of a given environment and its actors. On one hand, this is partly facilitated by the recent surge in representational capacity and flexibility of learned representations. On the other hand, the emergence of enhanced depth-sensing modalities such as high-quality stereo vision, monocular depth estimation, LiDAR, Radar offer new geometry-encoding cues, which are highly invariant with respect to view, appearance and photometric variations. These spatial cues, along with appearance attributes, are often exploited in robotic perception and interaction tasks, such as pose-aware grasping and path planning.

3D object pose denotes the spatial transform needed to align the coordinate reference of an observed object with that of the observer. As depth data contains distinctive cues linked to the sought translational and rotational object pose parameters, in this paper we present a focused study on examining the data quality of monocular and stereo depth modalities in light of a learned pose estimation task.

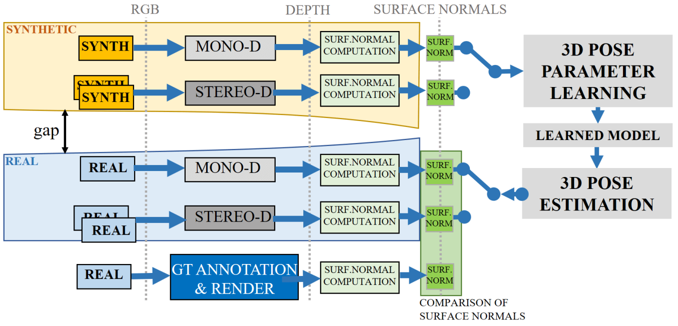

Fig. 1. Overview of the various depth and surface normals generation pipelines from synthetic and real data. Depth (computed surface normals) from synthetic images is used for training a pose-aware detector, which is tested on real images. Our proposed ground truth depth generation scheme is used to generate reference depth/surface normals data.

A primary motivation of our work stems from the fact that models trained on synthetic data often exhibit a severe degradation when facing the real data domain [8], or learning requires a photorealistic pipeline [2] to close the gap between simulated and real data. To mitigate this problem, we propose depth computation from synthetic images, with the objective to derive a representation exhibiting less synthetic qualities. Depth data estimated from synthetic images (via monocular or stereo estimation schemes), however, still might convey specific characteristics, which limit generalization towards real-world situations. Therefore, we propose a depth-data-specific comparison based on the computed surface normals to examine how quality discrepancies of different depth modalities relate to each other. Furthermore, we also examine the use of such training data with synthetic origins in the context of learning 3D pose-aware detectors, as it is described later on.

To support a quantitative comparison between the different depth modalities, we also propose a novel quantitative evaluation pipeline based on a simple ground truth generating procedure, yielding dense metric depth ground truth. Comparison of monocular depth estimates to a metric depth ground truth, however, is not straightforward due to the lack of metric scaling. To this end we propose an object-centered evaluation scheme, which compares computed surface normals at an object level and in a pixel-wise manner. Finally, to validate that a given transition to depth data narrows the synth-to-real gap, we present pose estimation experiments purely trained on depth from synthetic imagery. These experiments employ a baseline encoder-decoder-type pose estimation methodology and cylindrical objects as training and test objects. The presented data generation, evaluation and pose estimation scheme, however, is generic and also

*This work was partially carried out in the HOPPER project, supported by the “ICT of the Future” programme of the Austrian Research Promotion Agency (FFG)

[1] Assistive and Autonomous Systems, Center for Vision Automation and Control, AIT Austrian Institute of Technology, 1210 Vienna, Austria {andreas.kriegler, csaba.beleznai}@ait.ac.at

[2] Visual Computing and Human-Centered Technology, TU Wien Informatics, 1040 Vienna, Austria margrit.gelautz@tuwien.ac.at

applicable to more complex object geometries.

In summary, the paper proposes the following three contributions:

- Ground truth generation: we introduce a novel and generic annotation pipeline for computing dense and accurate depth and pose data for a wide variety of real scenes,
- Depth quality assessment: we propose a quantitative assessment scheme, comparing monocular and stereo-based estimates to ground truth via an object-centered analysis of computed surface normals,
- Initial results for 3D pose estimation trained on synthetic data: we demonstrate the feasibility of inferring 3D pose in real images via learned models trained on monocular and stereo depth normals, estimated from synthetic data.

The remainder of the paper is structured as follows: section II gives an overview on related work. Section III describes two data generation tasks: synthetic data generation for learning and depth ground truth generation for evaluation, both via Blender [3]. Section IV presents the proposed depth quality evaluation scheme. Finally, section V shows the applicability of a synthetic-data-based training pipeline to learn and predict object poses in real and synthetic images.

II. RELATED WORK

Recent research activities targeting learned representations of geometric traits encompass a large set of works, given that geometric shape and structure are intrinsic object properties which are highly invariant for different viewing and illumination conditions. This emerging field of geometric deep learning is well summarized in [4], [5], where geometric principles are highlighted to explain regularities often observed in the physical world, i.e. gravitational or right-angle structuring of man-made objects. Depth data naturally conveys geometric information, therefore understanding depth computation, its data characteristics and its failure modes are highly pertinent. [27] outlined four steps commonly encountered in classical stereo image pipelines. Despite representational advances via Deep Learning, these steps continue to play a key role [37]. 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 [36]. One of the first methods was [11] who also introduced scale-invariant evaluation metrics to measure the quality of the estimated depth maps. The proposed evaluation technique seeks an optimum depth scaling best aligning estimated depth and ground truth, a search step which is sensitive in presence of large depth discrepancies. Later works have explored continuously improving representations to learn a robust correlation between the appearance of a scene and its geometry [13], [22], [21]. Some works [20] approached this learning task as part of a multi-task learning scenario, where estimating the apparent motion and scene depth (from the viewpoint of a mobile observer) are formulated as two correlated and mutually-supporting learning tasks. An enhanced generalization of monocular depth models is attained via a mixture of datasets in [24]. Recent representational advances based on vision transformers [9], exploiting the attention-mechanism [32] are capable to accurately capture long-range semantic relations [23], see also [19] for a survey.

Nevertheless, the task of inferring absolute depth from a single image is an ill-posed problem, most prominently because of the prevailing scale ambiguity, making it unreliable in certain situations. These shortcomings lead to the conclusion in [28] that stereo vision is still required for accurate depth estimation, as stereo methods employ a principled, well understood, multi-view processing framework using concepts of the pinhole camera model [16]. In stereo vision, ambiguities are generated from other sources. In particular, stereo matching - that is, finding corresponding points in two (or more) stereo images for disparity estimation - is typically challenged by homogeneous image regions, repetitive patterns, depth discontinuities and occlusions, and particular surface reflectance properties. CNNs are known to be very powerful feature extractors and multiple learning-based deep neural network architectures exploit this capability for enhanced feature matching of stereo images, such as in [6], [12], [7]. AANet [34] provides a very good speed-accuracy trade-off.

While both monocular and stereo-based depth reconstruction methods have their individual advantages and limitations, one of the goals of our work is to provide a quantitative comparison of the two approaches in the context of a geometric deep learning task. Research works having a similar data characterization scope are still lacking. Although [28] provide a comparison between depths generated from monocular frameworks vs. stereo-setups, it is largely qualitative, and it does not use state of the art methods from the respective fields. [24] proposes several dataset-specific metrics, which are nevertheless difficult to relate across different datasets.

Finally, our paper is closely related to 3D object pose estimation from appearance and/or depth cues. Representational advances in recent years have resulted in an increasing accuracy and robustness with respect to clutter, occlusions and pose-ambiguous object types. This evolution is prominently reflected in the BOP Challenge series [18]. This paper leans on its resulting insights, that data availability and the domain gap between synthetic training and real test often represent a hurdle. These findings motivate us to devise data generation schemes which yield data conveying spatial cues and better bridge the domain gap.

III. GROUND TRUTH FOR DEPTH AND POSE

In the present data-driven era of Deep Learning, model performance is closely linked to the quantity and quality of training data [24]. The generation of synthetic data using GANs [14] has proven successful, while the inclusion of data from different sources such as YouTube videos [1] or movie datasets [24] is gaining interest as well. Nevertheless, the exploitation of frameworks commonly used in computer graphics applications is still largely unexplored [25]. One such program is Blender [3], which is commonly used to

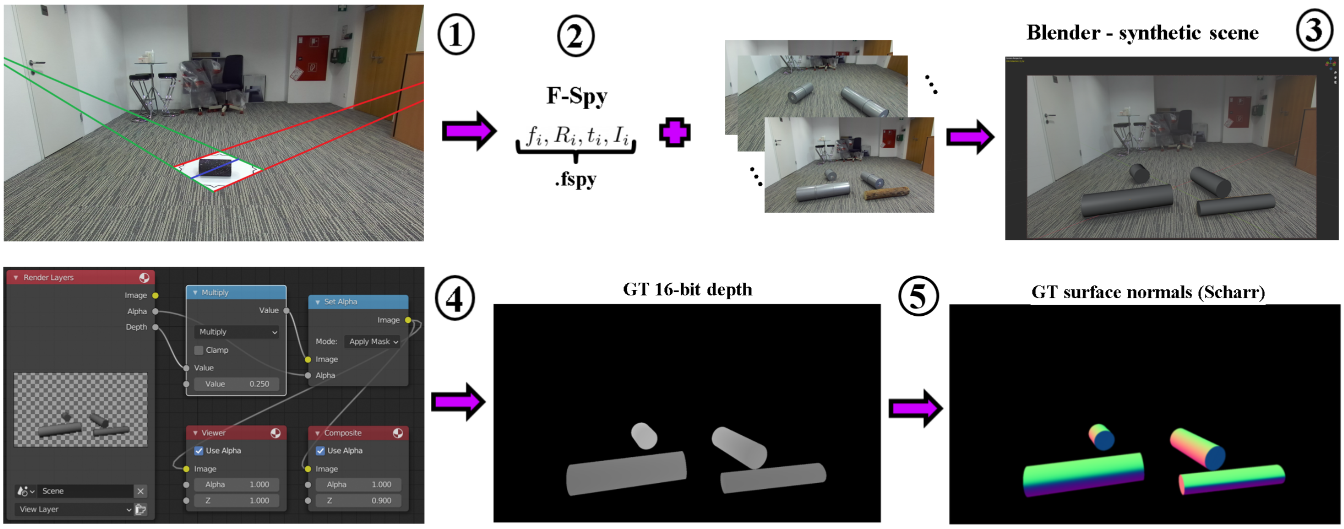

Fig. 2. Our proposed 3D object annotation pipeline yielding synthetically-correct per-pixel depth values and object 6DoF poses for real images captured with an RGB camera. The pipeline comprises five steps: 1) Alignment of two sets of parallel lines (red and green) where the sets are orthogonal, forming the ground plane. Additionally, a line segment of known length (blue) is set. 2) Creation of synthetic camera with estimated intrinsic and extrinsic parameters. 3) Camera and frames to be annotated are imported in Blender to create a synthetic twin of the scene. 4) Rendering of 16bit depth images. 5) Final transform of depth maps to surface normal images.

render synthetic imagery based on modelled or procedurally-generated scenes. In this work the use of the Blender platform is two-fold: training data generation for learning pose-aware object detectors, and 3D scene annotation for generating dense depth ground truth. These functionalities are explained in detail below:

Synthetic data generation: We procedurally generate a large synthetic dataset (consisting of cylindrical objects) along with 3D pose annotations. Diversity is introduced in form of various spatial object configurations and varying view parameters. We generate a rich set of object-specific annotations in form of 6DoF pose parameters, metric object dimensions, 2D bounding box, center point location and occlusion indicators using ray-tracing. Example renders can be seen at the top left side of Fig. 3, with textures randomized for each view. Please note that texture plays only a minor role as the RGB domain is not used directly for learning. Random textures, on one hand, generate a notion of the governing perspective and corresponding locations, thus facilitating the task of monocular and stereo depth estimation. Furthermore, random textures also introduce small-scale texture-induced monocular depth artifacts, thus robustifying a learned model with respect to locally corrupted depth/normal data. 56k of such synthetic samples (resolution 768×512px) are used to train and validate our object location and pose regression models, as described in Section V.

3D scene calibration, ground-truth-depth generation: Trained pose-aware models shall be evaluated on a real image dataset. To this end, we capture 120 rectified stereo image pairs using a Stereolabs ZED2 stereo camera [30] with a resolution of 1920×1080px. Captured images depict four different office environments (further on denoted as *scenarios*) with a variable number of cylindrical objects lying on a common ground plane. To generate a dense reconstruction for all scenes with as little effort as possible, we rely on a simple photogrammetric concept. We employ the technique by Guillou et al. [15], requiring two vanishing points and a line segment of known length. The two vanishing points can be easily defined by two line pairs, pairwise orthogonal to each other in the real world. Given these inputs, the camera rotational and translational parameters can be determined, along with its focal length. To perform the calibration and scene reconstruction, we execute following steps:

- Calibration: for a given scenario, we assume a stationary camera mounted on a tripod. We create a blank rectangular (cardboard) shape with at least one known dimension to use as a calibration target (see Step 1 in Fig.2). In the very first frame of a given scenario, the edges of the calibration target can be used to manually delineate two pairs of parallel line segments, yielding the sought camera parameters. In later images of a given scenario the target can be removed, since the camera views remain stationary. The publicly available fSpy toolkit [31] offers an interactive interface for the calibration algorithm [15]. It also provides a Blender camera generator functionality to create an equivalent camera within Blender, adequately oriented, translated and scaled, existing within a 3D space defined by the line segments leading to the vanishing points.
- 3D scene annotation in Blender: in this step we would like to spatially align a number of geometric objects within the 3D metric space of the camera. Each image of a scenario contains N ($1 < N < 4$) cylindrical objects, randomly placed in a lying pose. We measure the dimensions (length, radius) of these objects, in order to create cylindrical primitives of the same metric size in Blender. If objects are known to be on a common ground plane, object dimensions are not necessary.

Nevertheless, known dimensions significantly constrain the possible pose space where placed 3D objects are aligned with the apparent projections in the view seen through the camera (Steps 2 and 3 in Fig.2). This step also naturally leads to object pose attributes (orientation, translation) with respect to the camera.

- Dense ground-truth-depth generation: after aligning all 3D objects within Blender, the scaled Z-depth information can be rendered for the given scene. This depth information contains the metric depth for every scene point, similarly to a depth measurement via calibrated stereo cameras. This step is executed programmatically and produces float-valued depth entries for every image point where camera rays hit a previously placed object (see Step 4 in Fig.2). Note that a scaling factor is introduced for visualization only.

The presented calibration and ground truth generation scheme represents a straightforward annotation workflow for 3D object pose and depth data. It is applicable for a wide range of object geometries, scale ranges and arbitrary viewpoints. A detailed documentation, sample scene and code can be found at [*url*].

IV. EVALUATION OF MONOCULAR AND STEREO DEPTH DATA

In this section we describe a data-oriented evaluation methodology for three state-of-the-art depth estimation schemes. Our comparison targets the quality evaluation of structure-encoding surface normals (derived from the depth data), in the context of representation learning for 3D pose estimation. The three selected estimation schemes consist of two monocular depth estimation methods MiDaS_v2.1 [24] and MiDaS_v3.0 [23], and a stereo depth estimation model AANet [34]. Further on, MiDaS_v2.1, MiDaS_v3.0 (using the *DPT_large* model) and AANet are denoted as MiDaS, DPT and AANet for brevity.

MiDaS [24] is a CNN-based depth estimator using the framework of [33] with a ResNet [17] backbone. It was trained using up to ten different datasets, including 3D movies, leveraging the large data quantity and diversity for generalization. DPT [23] is a vision transformer trained for multiple dense prediction tasks including depth estimation. An argument for transformers is that they are able of capturing long-range semantic relationships in images [9], which should enforce stronger global structural consistency in the depth results; a trait which is often lacking for CNN-based monocular-depth frameworks [23]. Lastly, AANet [34] consists of an adaptive aggregation model for multi-scale disparity cost aggregation, resulting in an efficient stereo matching scheme. We employ the AANet *kitti*2015+ model which incorporates GANet [35] for feature matching.

A direct pixel-wise comparison between a monocular depth estimate and a ground truth is not straightforward, as monocular-depth frameworks generate disparity (inverse depth) values with no metric scaling. Furthermore, different monocular models yield disparity (depth) estimates with substantially different scaling factors. Therefore, common stereo vision evaluation metrics - assuming data living in a metric space - cannot be applied. To overcome this problem, we propose an object-centered scaling scheme performing a normalization within object-specific regions. The objective of this scaling step is to bring monocular, stereo and ground truth depth data within object foreground regions into a scale-normalized form. The input for this scaling is a depth image D_i, where $i = \{0,1,2\}$ indicates ground truth, monocular and stereo depth, respectively. A corresponding object foreground mask m_0 is also needed (generated via the ground truth generation process) to spatially constrain the set of pixels included into the normalization step. This mask contains unit entries at all object locations, denoted as object mask region m_0^{Obj}, and zeros elsewhere. The scaling operation is performed as:

$$D_i^* = D_i \,/\, max(1, D_i[m_0^{Obj}]), \tag{1}$$

resulting in D_i^*, a scaled depth image containing unit-normalized values within the object mask region. This subset of normalized depth values is used exclusively for further computation steps towards a quantitative comparison.

We adopt this object-foreground-based normalization procedure for ground truth, stereo and monocular depth data, resulting in depth values within the object foreground regions scaled to a common range. Instead of using the scaled depth values for an evaluation, we investigate its spatial derivatives, in form of surface normals. The choice of opting for surface normals stems from a representational consideration: when seeking to learn representations for objects situated at varying distances from a camera, computed surface normals exhibit less variation than depth data. We use a simple procedure to transform depth images to surface normals. First we calculate pixel intensity changes as derivatives d_x and d_y using the Sobel kernel [29]. With the intensity gradients we build local support planes, whose normal vectors can be seen as the normal vectors of the object surface in those pixels. As matrix norm we use the Frobenius norm. Since the Sobel kernel size is a configurable parameter, most commonly 3×3, 5×5 or 7×7, we examine resulting surface normal quality variations in this regard, and also include an alternative Scharr 3x3 kernel [26]. Images of computed surface normals are visualized by mapping the directional vector components to respective 8-bit RGB channels. The 3D vectors of surface normals are compared to ground truth in a pixel-wise manner using a 3D cosine similarity, yielding a similarity score in the range of $[-1,1]$. Vector similarity scores are mapped to a $[0,1]$ range and a cumulative score from all object foreground regions is formed.

V. RESULTS AND DISCUSSION

In this section, we present results on comparing depth data quality in terms of pixel-wise surface normal similarities with respect to a corresponding ground truth. The comparison is generated for a real dataset using the presented three (MiDaS, DPT and AANet) depth computation modalities. In addition to a quantitative evaluation, we also present qualitative results on the depth quality and experimental outcomes for 3D pose estimation. In the following, we describe our dataset and related evaluation results, followed by qualitative depth and pose estimation results.

Dataset and evaluation results: our 120 real-image dataset (see Section III) was captured using 4 distinct viewpoint setups, each scene containing 30 random object configurations. For each of the 120 images a corresponding depth ground truth was computed. Table I displays surface normal similarities computed from the MiDaS, DPT and AANet methods with respect to the ground truth. The table also shows the effect of the varying kernel-size used for surface normal computation. The kernel size of -1 denotes the Scharr, the other numbers relate to the Sobel kernel size. As it can be seen from the table, the two monocular depth estimation methods on an average produce very similar quality. While DPT tends to produce more geometric detail, in case of our targeted, smoothly varying surfaces it did not lead to enhanced scores. On the other hand, the AANet stereo matching scheme clearly outperforms the monocular models in all cases. The elongated cylindrical objects are long enough to call for the need of estimating accurate far-range structural correlations; a trait where monocular methods are still lacking behind the quality of the stereo depth data. Monocular methods generate a spatially-smooth output, where derivative kernels of increasing size deteriorate the captured geometric details. Therefore, a small kernel size of $3px$ seems to produce optimum results. AANet, on the other hand, benefits from larger kernels, suppressing the noise associated with the disparity estimation process.

Qualitative depth results: Fig. 3 displays a large set of qualitative results, partially generated for synthetic renders (Fig. 3 left half), partially for our real-image dataset. To facilitate the interpretation of depth quality, besides false-color depth images we also display two views of the point cloud representing the given scene. As it can be seen from the point cloud views, monocular depth estimation is relatively accurate when considering it locally, but at a large scale (especially near the image boundaries) significant spatial deviations occur. This observation also implies that, if pursuing an object detection or pose estimation task, local depth cues from monocular estimation can provide valuable hints, lifting many ambiguities associated with the monocular nature of the view. However, for problems requiring a global scale consistency, stereo pipelines still seem to be the more accurate and less data-dependent choice.

Qualitative pose estimation results: To examine the influence of data quality onto a 3D pose estimation learning task, we performed following experiments. Section III describes our data generation step for learning. The synthetic data, as single images or stereo pairs, are used within the respective monocular (MiDaS, DPT) and stereo (AANet) pipelines to generate depth data. The computed surface normals from depth and corresponding pose annotations {*class, 2D center, depth, angular parameters*} represent the input of our learning scheme. Given these inputs, an encoder-decoder-type framework (CenterNet [10]) was used to train depth-modality-specific models for 3D pose estimation (see also Fig. 1). We observe fast convergence within 3-5 epochs.

TABLE I

QUANTITATIVE COMPARISON OF SURFACE NORMALS COMPUTED USING DEPTH CUES FROM THREE MODELS. (HIGHER IS BETTER).

	k	*MiDAS_v2.1*[24]	*DPT_Large*[23]	*AANet*[34]
Scene 1	-1	0.7861	0.7898	0.8801
	3	0.7861	0.7898	0.8801
	5	0.7846	0.7887	0.8888
	7	0.7826	0.7870	0.8922
	avg.	0.7849	0.7888	**0.8853**
Scene 2	-1	0.8596	0.8722	0.9291
	3	0.8596	0.8722	0.9291
	5	0.8579	0.8714	0.9362
	7	0.8559	0.8702	0.9387
	avg.	0.8583	0.8715	**0.9333**
Scene 3	-1	0.8382	0.8005	0.8930
	3	0.8382	0.8005	0.8930
	5	0.8364	0.7984	0.9012
	7	0.8341	0.7958	0.9041
	avg.	0.8367	0.7988	**0.8978**
Scene 4	-1	0.8650	0.8604	0.9212
	3	0.8650	0.8604	0.9212
	5	0.8630	0.8588	0.9294
	7	0.8606	0.8567	0.9324
	avg.	0.8634	0.8591	**0.9261**

We demonstrate the applicability of our process to learn 6DoF pose parameters from purely synthetic data and perform prediction in real images. Fig. 3 shows pose estimation results (rows 6, 11 and 16) for the individual depth modalities. As it can be seen from the figure, in the synthetic domain results exhibit a high recall and a high pose estimation accuracy. In the real domain, however, inference on surface normals from monocular techniques shows several failure modes, a lower recall and precision, in form of occasionally hallucinating cylindrical objects within the nearby structural clutter. When using stereo-depth-based surface normals, however, results improve. Recall is still lacking, but no objects are hallucinated. Based on these results we believe that spatial cues derived from synthetic images can represent a way towards learning geometry-aware representations of objects and pose, which also exhibit validity within the real world.

VI. CONCLUSIONS

We present a geometrically-inspired depth data analysis scheme comparing surface normal cues from monocular and stereo-based pipelines, with object detection and 3D pose estimation tasks in mind. To support the data evaluation task, we propose a novel ground truth generation scheme, where dense depth and pose data can be created with little manual interaction. Our evaluations with respect to ground truth indicate that stereo-depth prevails in terms of data quality when compared to monocular depth, especially if a long-range depth data consistency is required. However, we demonstrate, that monocular depth still captures relevant local geometric details, which is sufficient to learn pose-aware object detectors from purely synthetic data. The demonstrated transition from the synthetic to the real domain seems to offer further geometry-aware analysis perspectives, while exploiting monocular or stereo depth cues.

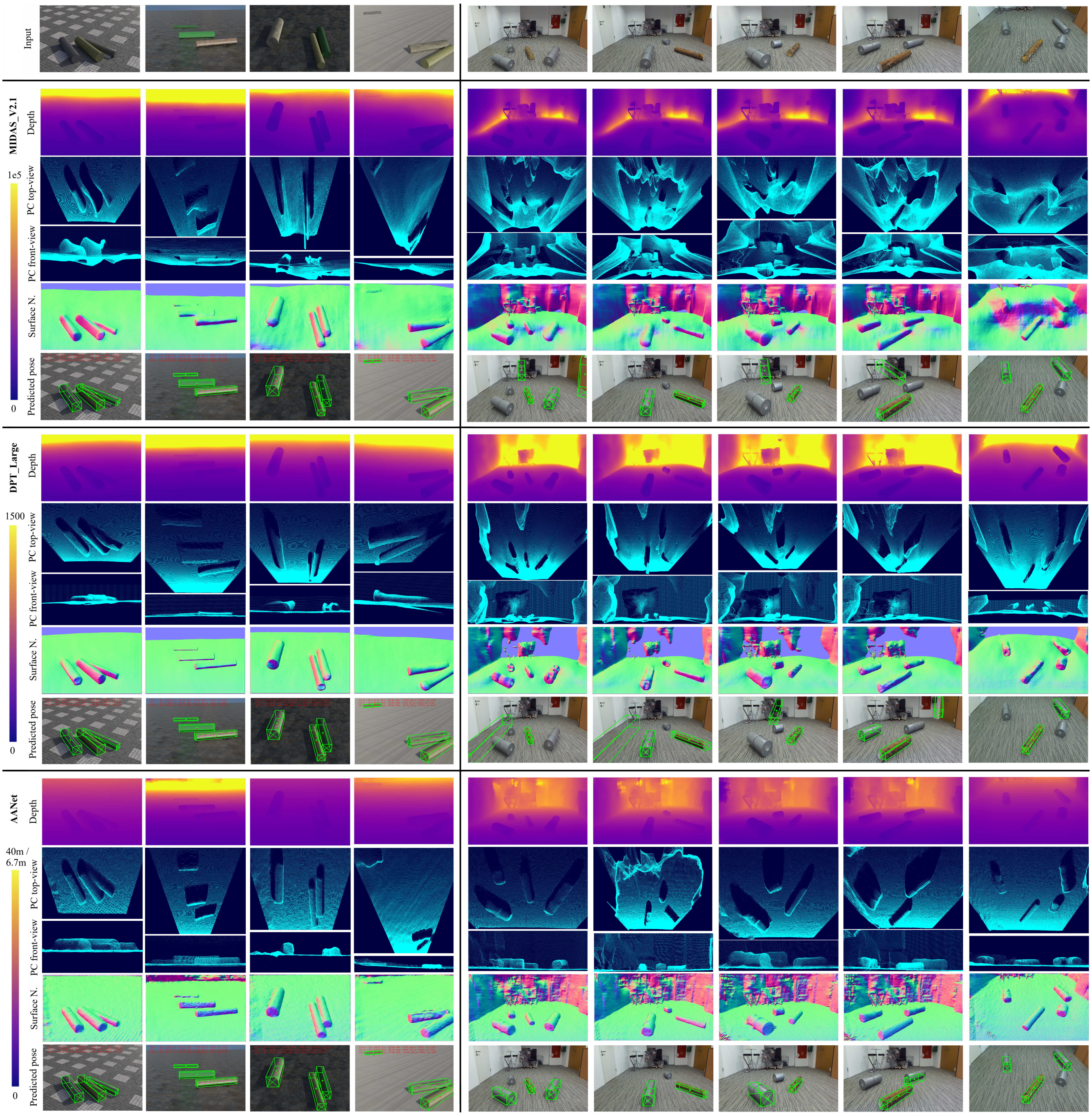

Fig. 3. For the synthetic data domain (left) and real images (right) we visualize example input images as well as depth images, depth point clouds and surface normals obtained using each of three depth estimation methods. The final rows for each model show CenterNet pose estimation results trained on surface normals from the large-scale synthetic dataset.

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, A. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *ArXiv*, vol. abs/1609.08675, 2016.

[2] H. Alhaija, S. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *International Journal of Computer Vision (IJCV), 2018*, 2018.

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

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

[5] 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.

[6] Z. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang, "A Deep Visual Correspondence Embedding Model for Stereo Matching Costs," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 972–980.

[7] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, H. Li, T. Drum-

mond, and Z. Ge, "Hierarchical neural architecture search for deep stereo matching," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 22 158–22 169.

[8] C. Doersch and A. Zisserman, "Sim2real transfer learning for 3d human pose estimation: motion to the rescue," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[9] 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.

[10] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint Triplets for Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 6569–6578.

[11] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems*, vol. 3, 2014, pp. 2366–2374.

[12] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "DeepStereo: Learning to Predict New Views from the World's Imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 842–857.

[13] R. Garg, B. G. Vijay Kumar, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9912 LNCS, no. April, pp. 740–756, 2016.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, p. 139–144, Oct. 2020.

[15] 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.

[16] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[18] T. Hodaň, M. Sundermeyer, B. Drost, Y. Labbé, E. Brachmann, F. Michel, C. Rother, and J. Matas, "BOP challenge 2020 on 6D object localization," *European Conference on Computer Vision Workshops (ECCVW)*, 2020.

[19] 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.

[20] C. Luo, Z. Yang, P. Wang, Y. Wang, W. Xu, R. Nevatia, and A. Yuille, "Every Pixel Counts ++: Joint Learning of Geometry and Motion with 3D Holistic Understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2624–2641, 2020.

[21] S. Mahendran, "Geometric Deep Learning for Monocular Object Orientation Estimation," Ph.D. dissertation, Hopkins University, 2018.

[22] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5667–5675, 2018.

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

[24] 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.

[25] S. Reitmann, L. Neumann, and B. Jung, "BLAINDER-A Blender AI Add-On for Generation of Semantically Labeled Depth-Sensing Data," *Sensors*, vol. 21, no. 6, 2021.

[26] H. Scharr, "Optimal filters for extended optical flow," in *IWCM*, 2004.

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

[28] N. Smolyanskiy, A. Kamenev, and S. Birchfield, "On the Importance of Stereo for Accurate Depth Estimation: An Efficient Semi-Supervised Deep Neural Network Approach," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 11 200–11 208, 2018.

[29] I. Sobel and G. Feldman, "Sobel - isotropic 3 x 3 image gradient operator," A Talk at the Stanford Artificial Project, 271-272, 1968.

[30] StereoLabs. (2019) Zedcam2. Stereo Labs. [Online]. Available: https://www.stereolabs.com/zed-2/

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

[32] 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.

[33] K. Xian, C. Shen, Z. Cao, H. Lu, Y. Xiao, R. Li, and Z. Luo, "Monocular relative depth perception with web stereo data supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[34] H. Xu and J. Zhang, "AANet: Adaptive Aggregation Network for Efficient Stereo Matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[35] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 185–194, 2019.

[36] 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.

[37] 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.