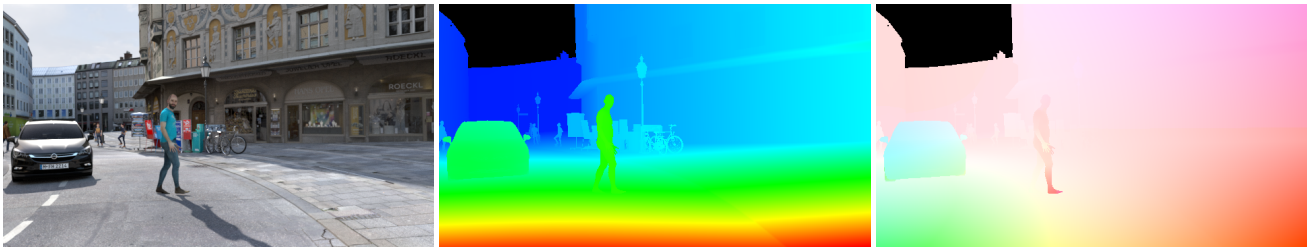
Automated Scene Flow Data Generation for Training and Verification

Extended Abstract

Oliver Wasenmüller
René Schuster
Didier Stricker
DFKI
firstname.lastname@dfki.de

Karl Leiss
Jürger Pfister
Oleksandra Ganus
Bit-TS
firstname.lastname@bit-ts.de

Julian Tatsch
Artem Savkin
Nikolas Brasch
BMW
firstname.lastname@bmw.de



(a) Image (b) Depth (c) Flow

Figure 1: We present a technology to automate the creation of synthetic images with dense scene flow ground truth.

ABSTRACT

Scene flow describes the 3D position as well as the 3D motion of each pixel in an image. Such algorithms are the basis for many state-of-the-art autonomous or automated driving functions. For verification and training large amounts of ground truth data is required, which is not available for real data. In this paper, we demonstrate a technology to create synthetic data with dense and precise scene flow ground truth.

1 INTRODUCTION

In the rapid development of autonomous driving functions and advanced driver assistance systems (ADAS), the motion estimation of all objects around the vehicle plays an essential role. Nowadays new cars are equipped with a multitude of sensors such as cameras, radar and ToF [17]. In order to determine the 3D motion of surrounding vehicles, pedestrians or unknown objects, so-called scene flow algorithms are utilized. Scene flow algorithms compute for each pixel of a stereo camera the 3D position as well as the 3D motion. This information can be used as an input for different driving functionalities.

Many state-of-the-art algorithms for computing scene flow geometrically have been presented in the Computer Vision literature [1, 7, 10, 11, 13, 15]. For their verification and evaluation dense ground truth data is required. Recently, there is a trend towards machine learning based flow estimation [5, 6, 9, 14]. These algorithms require even larger amounts of representative data for training and validation. While manual ground truth labeling on real data might be somewhat achievable for high-level annotations (e.g. 2D/3D bounding boxes, lane markings, etc.), precise labeling on pixel level is technically impossible.

Thus, we demonstrate in this paper a technology to create synthetic data with dense ground truth. After discussing state-of-the-art datasets in Section 2, we propose in Section 3 an approach the render scene flow. In Section 4 a technology for automated scene creation is presented followed by the data verification in Section 5.

2 RELATED WORK

As mentioned before, manual data labeling for scene flow is not possible. Even an expert can not determine for each pixel and with high precision its corresponding pixel in other images. There is also no hardware device, which can measure scene flow directly.

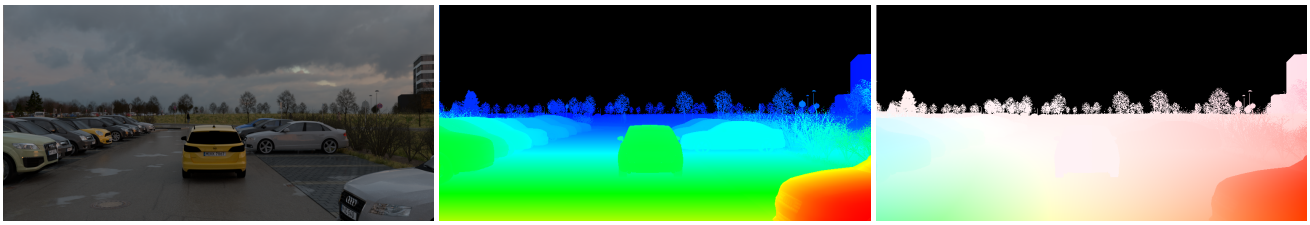
Some datasets, like the famous KITTI [4, 10], use a lidar with many scan lines in combination with a high-precision localization system. With this configuration it is possible to compute scene flow for static scene content. However, this flow is sparse and invalid for dynamic scene content.

Thus, we rely on synthetic data, where dense ground truth can be generated for any scene content. In state-of-the-art synthetic datasets, such as Sintel [2], virtualKITTI [3] or P4B [12], scene flow is not yet included.

3 SCENE FLOW RENDERING

Scene flow represents the 3D position as well as the 3D motion of each pixel in a reference frame of a stereo camera. In a synthetic

ACM Computer Science in Cars Symposium, 2018, Munich, Germany



(a) Image (b) Depth (c) Flow

Figure 2: Scene flow for (a) an image is illustrated as (b) depth image for the 3D positions and (c) flow image for the 3D motion.

scene it is straight forward to estimate the 3D position of each pixel with a simple back-projection. The more tricky part is the motion estimation, since the camera motion as well as intrinsic motions and deformations of the scene need to be considered.

First, we determine the 3D position of each pixel by projecting a ray through each pixel in the 2D image plane and check for the surface point hit of the ray. Besides the 3D coordinates of this hit, we store the corresponding triangle and its vertices. This is necessary to track the 3D coordinates also under scene motion and deformation. In case the scene contains intrinsic motion and/or deformation, the position of the triangle vertices will change in the next time step. However, since we stored their identifier, we can determine the position of the hit also after intrinsic motion and/or deformation. As long as the same ratio of distances to the vertices is maintained, the hit on a triangle can be determined precisely even after deformation. In a second step, the camera ego-motion needs to be considered. For that purpose, the hits after consideration of intrinsic motion and/or deformation are multiplied with the camera ego-motion. This gives us the final position of each pixel at the next time step with respect to the reference frame. The 3D motion of the scene flow are the 3D vectors between the starting hit and the computed hit. With this technology the scene flow can be determined even for pixels moving out of the image.

4 SCENE CREATION

The production of synthetic training and validation data has three main aspects to achieve for a long term impact: First the physical based sensor impression needs to be met. Second, the content and production of data must be verifiable and fulfill a quality assurance process. Third, the scalability of data production is essential for training, test coverage and validation; especially for machine learning algorithms. Today it is not proven which aspects of the real world are relevant for machine learning. Therefore, it is necessary to model a virtual sensor very close to the real sensor properties. Traditional content production workflows from game and film industry are based on thinking in whole scenes and using hand modeled content from artists. The artist can easily add additional meta-information into the scene so that labeling for a specific use case can be done automatically [8]. Unfortunately, in this approach the quality is highly depending on the artist and provides very low flexibility to adopt data to new requirements regarding sensor models as well as scene variation.

Our technology combines procedural and AI based generation of scenes on top of accurate maps including their semantic. Due to standardization, traffic signs and traffic lights are created according to their specifications. To organize assets, such as traffic signs, trees, vehicles, houses, etc., to reuse and automatically generate scene content in an efficient way, a database is built upon unique identifiers and a flexible data model. The automation in the scenario building process is established by the categorization of assets in the database in a modular clustering of geometry, materials and meta data. By linking meta data dynamically together along with the street semantic a scene and its variations are generated. Thereby boundary conditions such as weather, local and global illumination, traffic conditions, trajectories and injection of noisy data along their corner cases can be set as parameters. The pipeline scales with the number of hardware threads and is thus highly parallelized. Same scalability applies for a later rendering step where the sensor and ground truth outputs are processed.

5 DATA VERIFICATION

Although synthetic data allows to generate dense ground truth for scene flow, it is not guaranteed that the process is flawless. In order to validate the correctness of our data, we use automated sanity checks to verify its consistency. These checks include a round-trip check, forward-backward consistency, and ego-motion consistency. For a round-trip [16], we use the generated ground truth for motion and geometry to traverse two stereo image pairs in a cyclic order. From reference, to next time step, to stereo camera, back to the previous time step, and finally back to the reference camera. Up to subpixel errors through rounding, a round trip should always reach its starting pixel if the ground truth is consistent. For forward-backward consistency checks, the additional generation of scene flow in inverted temporal order is required, i.e motion from a reference time to the previously generated image. The corresponding backward motion of the target pixel of the forward motion should be the inverse. Having full control over the rendering process also allows to compute the motion of the virtual observer. By using ego-motion and ground truth depth, the correctness of the scene flow for static parts of the scene can be verified. These three tests help to generate accurate and correct data for training and evaluation.

6 CONCLUSION

In this paper, we presented a technology to automate the creation of synthetic scene content for dense accurate scene flow. Such data is required to train and verify different algorithms for driving functions in the context of autonomous driving and ADAS.

REFERENCES

[1] Aseem Behl, Omid Hosseini Jafari, Siva Karthik Mustikovela, Hassan Abu Alhaija, Carsten Rother, and Andreas Geiger. 2017. Bounding Boxes, Segmentations and Object Coordinates: How Important is Recognition for 3D Scene Flow Estimation in Autonomous Driving Scenarios?. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. 2012. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)*.

[3] A Gaidon, Q Wang, Y Cabon, and E Vig. 2016. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Fatma Güney and Andreas Geiger. 2016. Deep discrete flow. In *Asian Conference on Computer Vision (ACCV)*.

[6] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] Zhaoyang Lv, Chris Beall, Pablo F Alcantarilla, Fuxin Li, Zsolt Kira, and Frank Dellaert. 2016. A continuous optimization approach for efficient and accurate scene flow. In *European Conference on Computer Vision (ECCV)*.

[8] Nikolaus Mayer, Eddy Ilg, Philipp Fischer, Caner Hazirbas, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. 2018. What makes good synthetic training data for learning disparity and optical flow estimation? *International Journal of Computer Vision (IJCV)* (2018).

[9] Simon Meister, Junhwa Hur, and Stefan Roth. 2018. UnFlow: Unsupervised Learning of Optical Flow with a Bidirectional Census Loss. In *Conference on Artifical Intelligence (AAAI)*.

[10] Moritz Menze and Andreas Geiger. 2015. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[11] Zhile Ren, Deqing Sun, Jan Kautz, and Erik B. Sudderth. 2017. Cascaded Scene Flow Prediction using Semantic Segmentation. In *International Conference on 3D Vision (3DV)*.

[12] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. 2017. Playing for benchmarks. In *International conference on computer vision (ICCV)*.

[13] René Schuster, Oliver Wasenmüller, Georg Kuschk, Christian Bailer, and Didier Stricker. 2018. SceneFlowFields: Dense Interpolation of Sparse Scene Flow Correspondences. In *Winter Conference on Applications of Computer Vision (WACV)*.

[14] Ravi Kumar Thakur and Snehasis Mukherjee. 2018. SceneEDNet: A Deep Learning Approach for Scene Flow Estimation. *arXiv preprint arXiv:1807.03464* (2018).

[15] Christoph Vogel, Konrad Schindler, and Stefan Roth. 2015. 3D Scene Flow Estimation with a Piecewise Rigid Scene Model. *International Journal of Computer Vision (IJCV)* (2015).

[16] Oliver Wasenmüller, Bernd Krolla, Francesco Michielin, and Didier Stricker. 2014. Correspondence chaining for enhanced dense 3D reconstruction. In *International Conferences on Computer Graphics, Visualization and Computer Vision (WSCG)*.

[17] Tomonari Yoshida, Oliver Wasenmüller, and Didier Stricker. 2017. Time-of-Flight Sensor Depth Enhancement for Automotive Exhaust Gas. In *International Conference on Image Processing (ICIP)*.