KI ABSICHERUNG
Safe AI for Automated Driving

Lessons Learned on Synthetic Data Generation

Nicolas Gay[1], Maximiliano Cuevas[1], Ulrich Wurstbauer[1], Thomas Stauner[2], Oliver Grau[3], Korbinian Hagn[3], Falko Matern[4]

Synthetic Data Generation

Lessons Learned on Data Requirements

GL-DR-1: Define wether synthetic data will be used in combination with real data: use synthetic data to complement real data. This may lead to a smaller domain gap and faster convergence during training.

GL-DR-2: Prioritize data generation goals:
1) training: enough variance has to be included, 2) testing and validation: coverage of the ODD and corner cases, 3) validation of safety argumentation: oversampling of key aspects, e.g. number of pedestrians in a scene can be larger than in a real scene.

GL-DR-3: Match domain gap beween synthetic and real data to the tooling and budget availability. A larger domain gap is acceptable if it leads to e.g. increased detection robustness. For example, increasing content randomization at the expense of instance quality for robust detection of a large number of pedestrian poses.

GL-DR-4: Capturing object interactions, e.g. "scene contains pedestrians in the shadow of trees" is as important as the exact geometric properties of the scene. Both should be simultaneously defined by a scene authoring tool.

GL-DR-5 Scene specification languages like OpenScenario reduce ambiguity in the data generation specification. They allow for exchangability and reproducibility of results employing the use-case specific ontology.

GL-DR-6: Scenario randomization can easily generate object interactions which can be automatically filtered in a post-processing step before final data generation.

GL-DR-7: Scene design should cover performance limiting factors, such as object occlusion, lightning conditions, low object contrast, ambiguous or unusual object poses and locations, sensor effects, etc.

GL-DR-8: Crucial parameters and effects and their variations for the specific use-case must be identified in order to produce the right (minimal) amount of data. These parameters must be implementable by the tool-chain, e.g. HDRI (fig. 1).

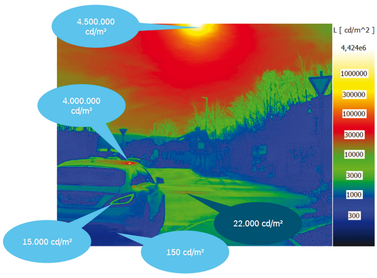

Figure 1: Challenge to reproduce High Dynamic Range in real world scenes in synthetic images

Lessons Learned on Data Production

GL-DP-1: Keep consistency of data labeling across the entire dataset in order to allow users to re-use data during incremental data production.

GL-DP-2: A data reader/loader must be available from the beginning of data production and must be consistent across data deliveries. This will ensure that data conforms to the specification and will also provide code-level compatibility in the downstream tooling.

GL-DP-3: A centralized big-data platform to manage and select data of interest for a particular use case allows for a streamlined development process. Alternatively, **a database with indexed metadata** rather than a full-fledged data management system will also allow the developer to select the data that he/she needs and thus save time and resources.

GL-DP-4: Combine use-case specific (self created) assets and generic asset libraries to provide the most cost-effective / time saving solution.

GL-DP-5: Define guidelines for the production of new assets in order to guarantee a consistent and for the use-case optimal asset quality **and provide these assets in a standardized format.**

GL-DP-6: Systematically identify and evaluate the sensitivity of production errors along the data generation toolchain in order to save post-production data correction time and effort and also provide better ML algorithm performance during training and testing

Lessons Learned on Data Analysis

GL-DA-1: Production toolchain should produce physically plausible results regarding all relevant effects of the use-case, e.g., the used material models support all relevant effects, and the render-tooling produces physical plausible irradiance values

GL-DA-2: Metadata from production tooling and post-processing can provide valuable insights regarding data quality, coverage and useability. E.g. correlation and analysis of parameters such as pedestrian positions (fig. 2) and poses, contrast measurements, etc., can can be used to compare datasets and correlate data properties to detection performance.

GL-DA-3: Human understandable features can be used as feedback to data production to iteratively improve simulation and data-content for the defined use-case. E.g. enriched metadata annotations describing scene, image and object properties (fig. 3).

GL-DA-4: Photorealistic data is not always required. The render engine should be capable to emulate physical effects of the sensor if needed but it should also be able to turn effects off in order to speed up computation.

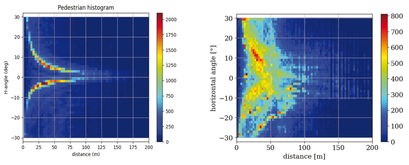

Figure 2: Pedestrian distribution according to camera's FOV derived from 3D Bounding Box data

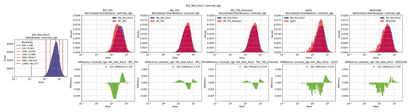

Figure 3: Comparing contrast (human understandable feature) distribution across KIA tranches and with external datasets as feedback to data production

Lessons Learned on Transfer to Real-World Application

GL-TA-1: On domain adaptation with insufficient target domain data: As long as a considerable amount of data from the target domain is available, for semantic segmentation models not a single approach in domain adaptation leads to significant performance improvements over using fine-tuning, but merely to improved convergence speed during training.

GL-TA-2: Limiting the domain (ODD) of a DNN model leads to better results than expanding the ODD and training over a larger dataset

GL-TA-3: Evaluating data from a real dataset on a DNN trained on synthetic data provides valuable insights on 1) dataset coverage, e.g. similarity of class distributions and , 2) the domain gap between both datasets. For instance adding an Error Generator (EG) provides an effective mean to reduce the domain gap as seen in Table 1 for KIA-Tr3 with and without Error Generator data.

Model trained on ...	mIoU Evaluation on ... (%)			
	Real World Dataset		Synthetic Dataset	
	Cityscapes	BDD100K	KIA-Tr4	KIA-Tr7
Cityscapes	**81.56**	58.16	52.11	63.67
BDD100K	66.06	**67.78**	53.75	55.10
KIA-Tr4	33.55	28.8	**77.6**	59.89
KIA-Tr7	49.83	36.35	58.22	**87.64**
KIA-Tr3	40.37	35.42	**66.95**	57.13
KIA-Tr3 + EG	**47.20**	**40.91**	66.13	**57.4**

Table 1: Comparison of mIoU on real and synthetic datasets for DeepLabV3+ (Semantic Segmentation)

Luxoft[1] A DXC Technology Company · BMW GROUP · MINI[2] · intel[3] · BOSCH[4]

For more information contact:
ngay2@dxc.com
l.cuevaseiroa@dxc.com

KI Absicherung is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Climate Action.

KI FAMILIE

Supported by:
Federal Ministry for Economic Affairs and Climate Action
on the basis of a decision by the German Bundestag

https://www.ki-absicherung-projekt.de · @KI_Familie · KI Familie

Image sources: Robert Bosch GmbH