KIDELTA LEARNING
Scalable AI for Automated Driving

Is In-Domain Data Required to Learn Multi-View Depth Estimation?

Philipp Schröppel, Artemij Amiranashvili, Thomas Brox

WP 4.1

Environment

Use-Case in Delta Learning

Robust depth estimation is an important component of autonomous driving and erroneous estimates can lead to fatal failures. Therefore, it is crucial that depth estimation works robustly in all possible scenarios.
The so-called open-world problem describes that it is not possible to cover all possible scenarios by the training data. To circumvent this, we demonstrate that a model for multi-view depth estimation can be trained only on randomized synthetic data and still generalize across domains. This reduces the risk of problems due to potential out-of-distribution samples in practical applications.

Technical Problem

Depth from images can be derived from multiple cues, e.g. single-view priors, or the motion parallax. Single-view priors are domain-specific, whereas the principle of motion parallax is generic and works across domains. Learned multi-view depth estimation models could exploit both cues. However, such models are typically trained and tested on in-domain data, which makes it difficult to judge their generalization capabilities. We therefore evaluate such models across different domains and ask the question if in-domain data is required to learn multi-view depth estimation.

Technical Solution

We extend the existing DispNet architecture to a multi-view stereo setting, as shown in Fig1. As training dataset, we render StaticThings, a static version of the synthetic FlyingThings3D dataset. To prevent our model from overfitting to the depth distribution of StaticThings3D, we introduce scale augmentation: GT translations and depths are scaled randomly during training.

Evaluation

We define test sets from diverse datasets: KITTI, ScanNet, ETH3D, StaticThings, RealThings. We evaluate in a zero-shot cross-dataset fashion without any fine-tuning. We use the Absolute Relative Error metric and report results in Tab 1.

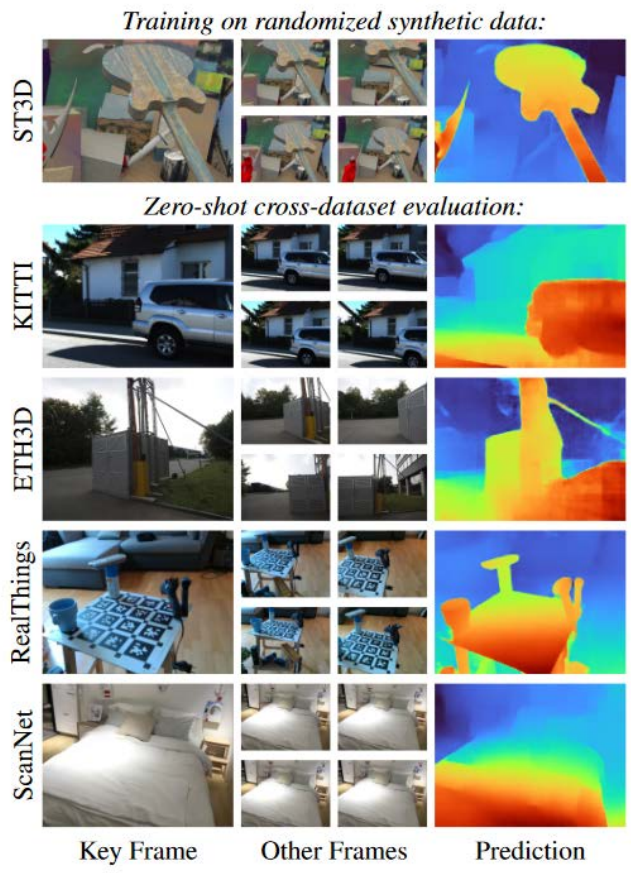

Figure 2: Qualitative results of our model on datasets from different domains.

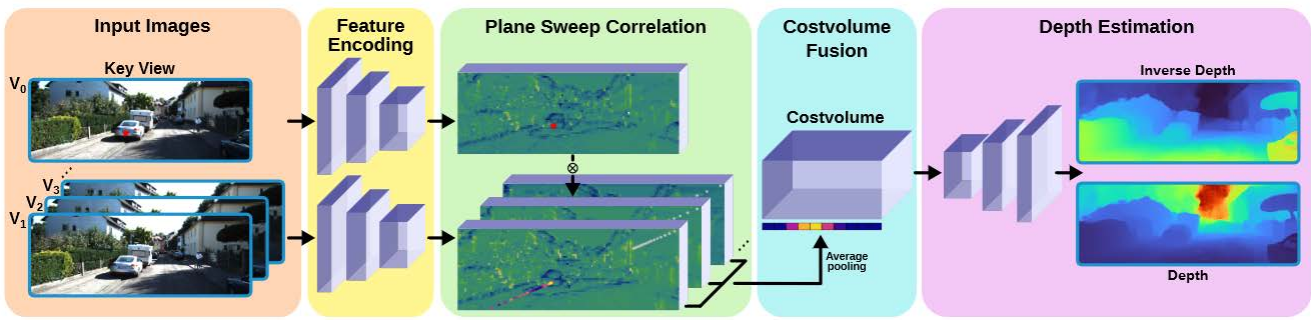

Figure 1: Our model is based on a DispNet architecture, but a) correlates views in a plane sweep stereo fashion, b) aggregates multi-view information by averaging costvolumes, and c) predicts inverse depth maps.

Approach	GT Poses	Output Scaling	KITTI	ETH3D	ScanNet	RealThings	StaticThings3D	Average
Median	✗	✓	0.41	0.32	0.23	0.28	0.56	0.36
DeMoN	✗	✓	0.16	**0.17**	0.12	0.15	0.37	0.19
DeepV2D (ScanNet)	✗	✓	0.18	0.19	**(0.06)**	0.12	0.34	0.18
DeepV2D (KITTI)	✗	✓	**(0.03)**	0.28	0.23	0.30	0.55	0.28
MVSNet	✓	✗	1.91	**0.14**	0.11	1.42	**0.23**	0.76
Ours	✓	✗	**0.13**	0.25	**0.09**	**0.06**	**(0.09)**	**0.13**

Table 1: Quantitative results of zero-shot cross-dataset evaluation, reporting the Absolute Relative Error (AbsRel).

For more information contact:
schroepp@cs.uni-freiburg.de

Partners

External partners

KI Delta Learning 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 Energy.

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