Semi-Supervised Disparity Estimation with Deep Feature Reconstruction

Julia Guerrero-Viu* Sergio Izquierdo* Philipp Schröppel Thomas Brox

guerrero, izquierd, schroepp, brox@cs.uni-freiburg.de

LINK TO POSTER

Women in Computer Vision

Problem Statement

- **Disparity estimation models trained on synthetic data have limited generalization**. Obtaining labels for supervised fine-tuning on the target domain is expensive.
- **Self-supervision based on view reconstruction** allows label-free training on the target domain, but **performs worse**, especially on challenging areas **due to limitations of the common photometric consistency**.
- **Self-supervision based on deep feature reconstruction** may help to overcome photoconsistency limitations. However, this **requires further analysis**.

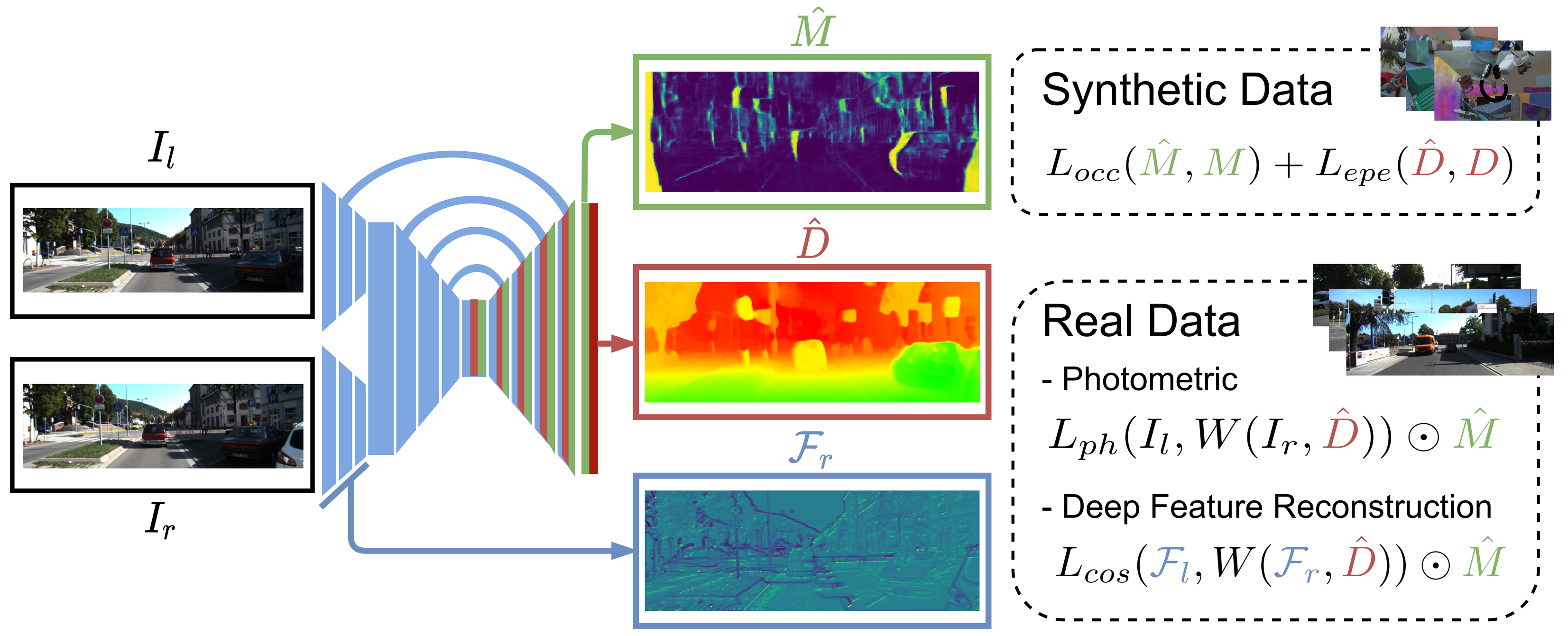
Contributions

- We propose a **semi-supervised pipeline for disparity estimation** with **supervised training on labeled synthetic data** and **self-supervised training on unlabeled real data**.

  → **Improves cross-domain generalization**.
- We perform a **thorough analysis of deep feature reconstruction**.

  → Shows the potential of deep feature reconstruction and **analyses problems that limit its effectiveness**.

Approach Overview



- We use a **DispNet architecture** [5] that predicts **occlusion masks** $\hat{M}$ and **disparity maps** $\hat{D}$ at multiple scales.
- We use a **supervised loss** for disparity and occlusion predictions **on synthetic data** and a **self-supervised reconstruction-based loss** for disparity predictions **on real data**.
- **For self-supervised training**, we experiment with **either photometric or deep feature reconstruction as supervisory signal**.
- We consider feature maps $\mathcal{F}_l, \mathcal{F}_r$ from the DispNet encoder (first three conv layers) for the deep feature reconstruction.

Semi-Supervised Pipeline

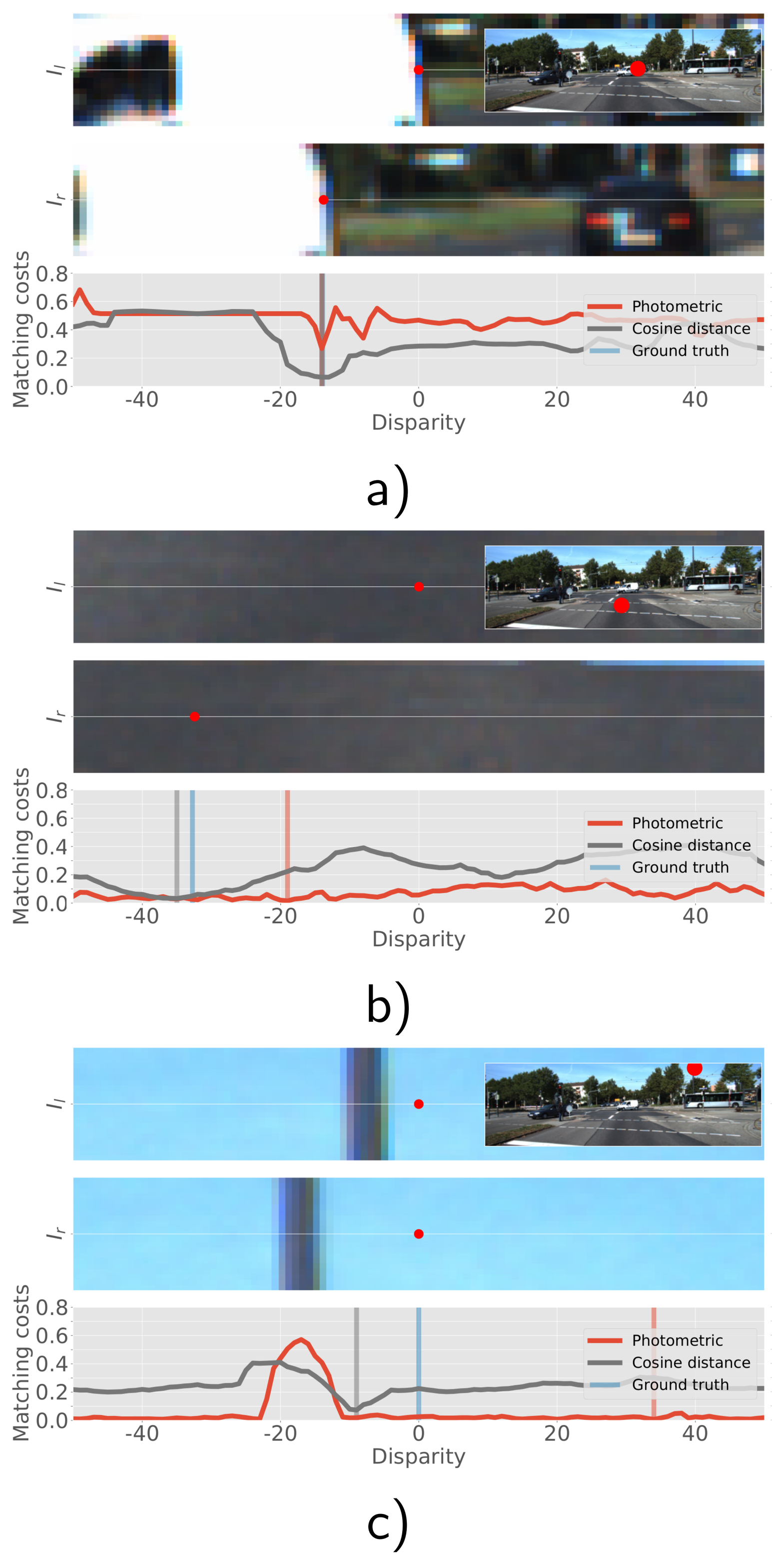
- Previous works either apply
  a) supervised pre-training on synthetic data and supervised fine-tuning on real world target domain data (e.g. [5]), or
  b) self-supervised training from scratch directly on the target domain (e.g. [3]).
- **We train from scratch in a semi-supervised fashion**: alternating batches of synthetic data (with labels) and real data (without labels).
- We use predicted occlusions to mask the loss on real world samples.

Deep Feature Reconstruction (DFR)

- Most works apply self-supervision based on view reconstruction: warp the right image according to the predicted disparities and measure reconstruction via photometric consistency.
- Instead, **we apply DFR**: warp and compute loss on feature maps.

  → **loss is based on consistency of the warped right image feature map and the respective original left image feature map**.

Analysis



We identity the following problems of DFR:

1. Higher sensitivity to occlusions.
2. Large dependence on the distance metric and resampling strategy.
3. Tainted information around disparity discontinuities due to convolutional aggregation.
4. Higher entropy on matching curves.
5. High gradient locality that complicates optimization.

On the left figure we illustrate (3) and (4). Despite DFR's clearly better response in texture-less areas (road), it fails near disparity discontinuities (sky-pole). Due to its higher entropy curve, DFR is less precise on object boundaries (van).

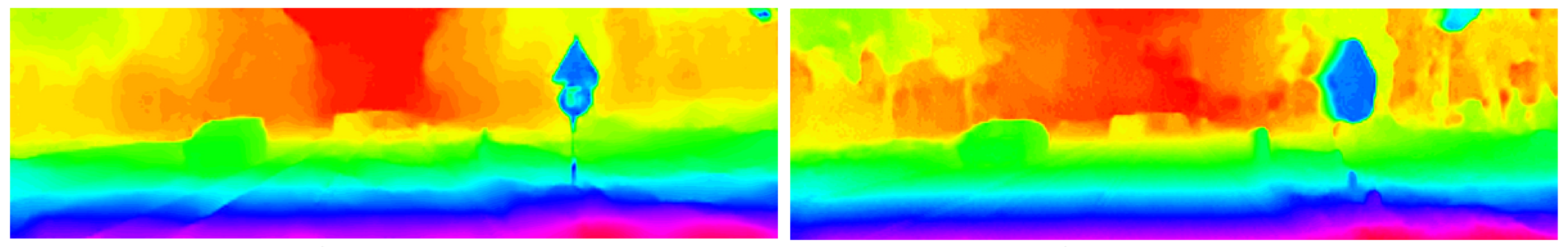
Experiments

- Training data: FlyingThings3D (synthetic, labeled), KITTI RAW (real, unlabeled)
- Test data: FlyingThings3D-Test, KITTI2015, ETH3D, Middlebury

Model	train DS	time	Endpoint Error (EPE) FT	K15	ETH3D	MidH
DispNet Supervised	FT	0.04	**1.69**	1.46	0.92	3.21
DispNet Supervised ft	FT + K15	0.04	3.05	(0.69)	1.99	3.79
DispNet SemiSup. PH	FT + (K)	0.04	1.77	**1.23**	**0.61**	**2.92**
DispNet SemiSup. DFR	FT + (K)	0.04	1.77	1.32	0.67	2.94
GWCNet-gc [4]	FT	0.32	1.65	2.35	1.73	5.08
GWCNet-gc ft [4]	FT + K12	0.32	5.63	**0.82**	1.09	5.41
LEA Stereo [2]	FT	0.30	**1.58**	1.98	0.87	4.72
Reversing PSMNet [1]	(K)	0.41	6.03	1.01	**0.51**	6.02

Results on FT 'cleanpass' test set and K15, ETH3D, MidH train sets. Train datasets are in brackets when no labels are used. We report all SOTA results by evaluating their publicly available models. We do not filter disparities>192.

Qualitative results on Kitti 2015:



a) Supervised b) Semi-Supervised

- Our results show **improved generalization** across domains, outperforming previous works in this setting.
- Our network is **drastically faster** than most SOTA models.

References

[1] Aleotti et al. Reversing the cycle: self-supervised deep stereo through enhanced monocular distillation. In *ECCV*, 2020.
[2] Cheng et al. Hierarchical neural architecture search for deep stereo matching. *NIPS*, 2020.
[3] Godard et al. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.
[4] Guo et al. Group-wise correlation stereo network. In *CVPR*, 2019.
[5] Mayer et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016.

Acknowledgements

We acknowledge partial funding by 'la Caixa' Foundation (LCF/BQ/EU19/11710058), by the German Research Foundation (BR 3815/10-1) and by the German Federal Ministry for Economic Affairs and Energy within the project "KI Delta Learning".