Semi-Supervised Disparity Estimation with Deep Feature Reconstruction

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

University of Freiburg

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

Abstract

Despite the success of deep learning in disparity estimation, the domain generalization gap remains an issue. We propose a semi-supervised pipeline that successfully adapts DispNet to a real-world domain by joint supervised training on labeled synthetic data and self-supervised training on unlabeled real data. Furthermore, accounting for the limitations of the widely-used photometric loss, we analyze the impact of deep feature reconstruction as a promising supervisory signal for disparity estimation.

1. Introduction

From 3D reconstruction to autonomous driving, many applications require depth estimates, which can be accurately obtained by predicting disparity from stereo images.

Inspired by the matching cost and aggregation ideas from traditional stereo algorithms, recent works have designed successful deep learning architectures that are trained for disparity estimation on labeled data [10, 9, 5, 2, 15]. However, most of these works report only fine-tuned results for each individual dataset, neglecting generalization across domains [9, 5, 2]. Trying to improve this, and given the cost of acquiring ground truth data on real scenarios, some works have adopted a self-supervised approach [20, 17, 19, 1]. They replace labels by a view reconstruction objective that maximizes consistency between the target image and the second image warped with the predicted disparity. However, self-supervised methods are still less precise, especially on fine details and challenging areas, partially due to the limitations of this photometric consistency [8]. Aiming to overcome these limitations, [18, 14] and [7] have explored a reconstruction error based on deep features, for monocular depth and optical flow. Nevertheless, performance only improved marginally after combining it with a photometric loss, which shows the need for further analysis.

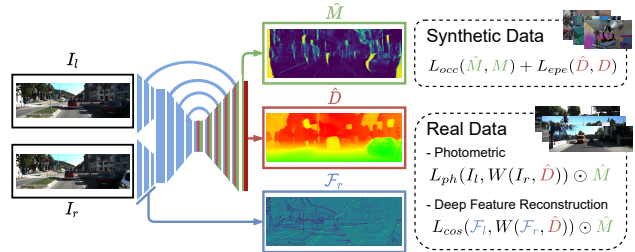In this work, we present two contributions that address the described problems: (1) we propose a semi-supervised

Figure 1. **Semi-supervised training pipeline**: We use supervised training on synthetic data and self-supervised training on real data samples, either with photometric or deep feature reconstruction.

pipeline for disparity estimation that improves cross-domain generalization by exploiting cheap synthetic labels and unlabeled data from real scenarios, and (2) we summarize the result of a thorough analysis of deep feature reconstruction (DFR) as consistency measure in self-supervised training, where we show examples of its potential and shed light on key reasons that limit its current effectiveness.

2. Semi-supervised pipeline

Aiming to improve the domain adaptation, we propose a general semi-supervised pipeline for disparity estimation, depicted in Figure 1. We train a single end-to-end network by alternating batches of synthetic and real data. Synthetic batches, whose labels are already available at no extra cost, are used to train in a supervised manner. On the other hand, we train with real-data batches in a self-supervised way, using either the well-known photometric [4] or deep feature reconstruction loss, as detailed in Section 3. Additionally, following [6], we train the network during supervised updates to predict occlusions. In real-data updates, we binarize the predicted occlusions and use them to mask out occluded areas in the self-supervised loss. Finally, we add synthetic occlusions [16, 15] to real-world data, which allows the network to learn about occlusions also from unlabeled samples. In this work, we use the DispNet-C [10] architecture, for its simplicity and efficiency, but our proposed pipeline is compatible with other disparity estimation networks.
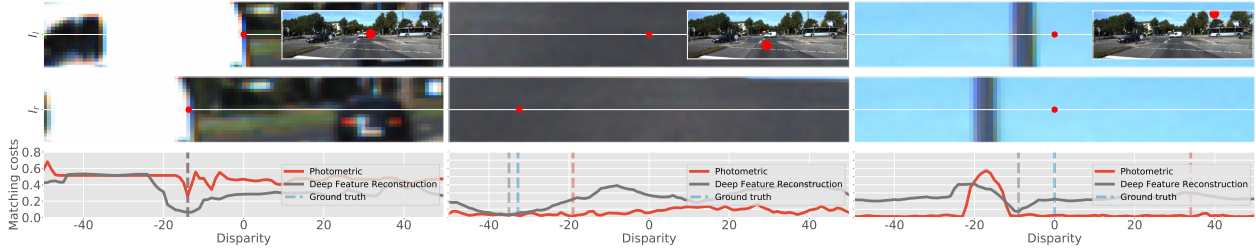
---

*Equal contribution

Figure 2. **Matching costs curves of DFR and photometric on three different scenarios**. From top to bottom: zoom-in of the left image (target point in red); same zoom-in on the right image (matching ground truth in red); obtained matching costs at different disparities along the epipolar line, aligned with the right image (minimum of the curves and ground truth are marked with vertical dashed lines).

3. Deep feature reconstruction

We investigate on DFR for self-supervision by substituting the RGB images with features extracted from the first three layers of DispNet. Due to their higher-level representation, deep features should provide a more robust signal on challenging areas, like texture-less or non-Lambertian surfaces. Given the predictions and feature maps at multiple scales, we resample each disparity map with nearest-neighbor interpolation to match all feature map sizes. We compute the dissimilarity between the left and the warped right feature maps using cosine distance, which resembles the computed dot product from the correlation layer.

4. Experimental results

We present results of our semi-supervised framework, using photometric loss (PH) and deep feature reconstruction (DFR). We use FlyingThings3D (FT) [10] for supervised updates, and KITTI Raw (K) [3] for self-supervised ones, equally balancing their number of samples per epoch.

Semi-supervised learning with photometric consistency: In Table 1, we compare our results with two supervised baselines. On KITTI2015 (K15) [11], the semi-supervised approach with photometric loss outperforms the supervised DispNet trained only on FT (-15.8% EPE), showing the improved adaptation to real-world domains. We even achieve this while increasing the error on the synthetic FT domain only marginally (+4.7% EPE), as opposed to the drastically increased error with supervised fine-tuning on K15 (+80.5% EPE). We also test the generalization to unseen environments using ETH3D [13] and Middlebury at half resolution (MidH) [12]. Remarkably, our semi-supervised approach generalizes also to these datasets, whereas fine-tuning on K15 only learns priors from KITTI but does not generalize to other real-world data.

Deep feature reconstruction analysis: As shown in Table 1, we obtain inferior results with DFR than with its photometric counterpart. Instead of directly combining DFR with a photometric loss as in previous works, we conducted an in-depth analysis of the reasons that currently prevent the

			Endpoint Error (EPE)			
Model	train DS	time	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 [5]	FT	0.32	1.65	2.35	1.73	5.08
GWCNet-gc ft [5]	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

Table 1. 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. Notice we do not filter disparities>192.

effectiveness of DFR. We identified the following issues: (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, and (5) high gradient locality that complicates optimization. We illustrate (3) and (4) on Figure 2. 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).

Comparison to the state of the art: Finally, we compare to state-of-the-art approaches, namely the supervised GWCNet [5] and LEAStereo [2], and the self-supervised Reversing PSMNet [1]. Results demonstrate the generalization problems of current supervised works. In contrast to this, our pipeline achieves reasonable results on FT and K15 and generalizes better across domains.

5. Summary

We have presented a novel semi-supervised pipeline and analyzed the influence of deep feature reconstruction for disparity estimation. Our results show improved generalization across domains, outperforming previous works in this setting. Based on our detailed study of DFR, we aim to exploit its potential in future work.

Acknowledgment

References

[1] Filippo Aleotti, Fabio Tosi, Li Zhang, Matteo Poggi, and Stefano Mattoccia. Reversing the cycle: self-supervised deep stereo through enhanced monocular distillation. In *16th European Conference on Computer Vision (ECCV)*. Springer, 2020. 1, 2

[2] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 2

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 2

[4] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3828–3838, 2019. 1

[5] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3273–3282, 2019. 1, 2

[6] E. Ilg, T. Saikia, M. Keuper, and T. Brox. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In *European Conference on Computer Vision (ECCV)*, 2018. 1

[7] Woobin Im, Tae-Kyun Kim, and Sung-Eui Yoon. Unsupervised learning of optical flow with deep feature similarity. In *European Conference on Computer Vision*, pages 172–188. Springer, 2020. 1

[8] Rico Jonschkowski, Austin Stone, Jon Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1

[9] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017. 1

[10] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 1, 2

[11] Moritz Menze, Christian Heipke, and Andreas Geiger. Object scene flow. *ISPRS Journal of Photogrammetry and Remote Sensing (JPRS)*, 2018. 2

[12] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nesic, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *GCPR*, volume 8753, pages 31–42. Springer, 2014. 2

[13] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[14] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conference on Computer Vision*, pages 572–588. Springer, 2020. 1

[15] Vladimir Tankovich, Christian Häne, Sean Fanello, Yinda Zhang, Shahram Izadi, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. *arXiv preprint arXiv:2007.12140*, 2020. 1

[16] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2019. 1

[17] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 636–651, 2018. 1

[18] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[19] Junming Zhang, Katherine A Skinner, Ram Vasudevan, and Matthew Johnson-Roberson. Dispsegnet: Leveraging semantics for end-to-end learning of disparity estimation from stereo imagery. *IEEE Robotics and Automation Letters*, 4(2):1162–1169, 2019. 1

[20] Chao Zhou, Hong Zhang, Xiaoyong Shen, and Jiaya Jia. Unsupervised learning of stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1567–1575, 2017. 1