Semi-Supervised Disparity Estimation with Deep Feature Reconstruction

Julia Guerrero-Viu\* Sergio Izquierdo\* Philipp Schröppel Thomas Brox
University of Freiburg

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 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
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 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 LEA Stereo [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

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 – Development of methods and tools for the efficient expansion and transformation of existing AI modules of autonomous vehicles to new domains”.

References