SELF-SUPERFLOW: SELF-SUPERVISED SCENE FLOW PREDICTION IN STEREO SEQUENCES

Katharina Bendig[1], *René Schuster*[2], *Didier Stricker*[1,2]

[1]Technische Universität Kaiserslautern
[2]DFKI – German Research Center for Artificial Intelligence
firstname.lastname@dfki.de

ABSTRACT

In recent years, deep neural networks showed their exceeding capabilities in addressing many computer vision tasks including scene flow prediction. However, most of the advances are dependent on the availability of a vast amount of dense per pixel ground truth annotations, which are very difficult to obtain for real life scenarios. Therefore, synthetic data is often relied upon for supervision, resulting in a representation gap between the training and test data. Even though a great quantity of unlabeled real world data is available, there is a huge lack in self-supervised methods for scene flow prediction. Hence, we explore the extension of a self-supervised loss based on the Census transform and occlusion-aware bidirectional displacements for the problem of scene flow prediction. Regarding the KITTI scene flow benchmark, our method outperforms the corresponding supervised pre-training of the same network and shows improved generalization capabilities while achieving much faster convergence.

Index Terms— Scene flow, Self-supervision, Occlusion, Stereo

1. INTRODUCTION

The safe navigation of robots and autonomous vehicles strongly depends on the sufficient perception of their surroundings. This especially necessitates the correct estimation of the motion of other traffic participants. Scene flow is one way to represent the rich perceptual information needed as it depicts the complete 3D motion field of objects in the environment and does not introduce additional ambiguity compared to its projection on the image plane (optical flow).

A currently famous approach for scene flow prediction is the usage of convolutional neural networks since they are able to outperform many traditional methods [1]. These methods are trained in a supervised manner and thus rely heavily on the availability of a large amount of ground truth data, which is difficult to obtain in a real world setting, especially for the problem of scene flow prediction. For that reason, synthetically generated images are often used for pre-training, since they allow an easy generation of dense ground truth scene flow. However, synthetic images display intrinsic differences to the real world images introducing an inherent representation gap between the training and test data sets. Real world data is often only utilized for an optional fine-tuning stage after the synthetic pre-training. This results in a reduced generalization for real world scenarios and thus makes the application in areas like autonomous driving challenging.

One way to address these obstacles is the utilization of self-supervision, i.e. the network obtains its training signal (loss) based on the available input data. However, there is currently a great lack in self-supervised approaches for stereo scene flow prediction. Therefore, we propose a novel method to solve scene flow prediction in this self-supervised setting for the first time. Our innovative strategy infers multiple displacements based on the availability of rich information due to the two stereo image pairs used. These displacements can be applied for occlusion reasoning as well as bidirectional reconstruction losses for arbitrary reference images. Our approach further utilizes a ternary census loss following the example of [2] and explores the effect of optimization techniques for self-supervision like cost volume normalization introduced by [3]. Our novel self-supervised training method for stereo scene flow prediction is called Self-SuperFlow. On the KITTI evaluation data set [4, 5], it vastly outperforms the equivalent supervised pre-training of the same network with 20 percentage points less outliers, while drastically cutting down the convergence time of the training.

2. RELATED WORK

2.1. Scene Flow Prediction

Optical flow approaches build the foundation for many scene flow prediction methods. Thus traditional approaches are often based on variational methods, which involve the minimization of an energy function [6, 7]. However, due to their iterative nature, long run time, and constancy assumptions, variational methods are non applicable in real world scenar-

This work was partially funded by the Federal Ministry of Education and Research Germany under the project DECODE (01IW21001).

ios. Another strategy is to assume rigidity, like e.g. in [5, 8], albeit objects in real world settings often do not move rigidly thus breaking the assumption. Therefore, the preferred method nowadays is the usage of neural networks, which are trained in a supervised manner [1, 9]. However, this depends strongly on the availability of a large amount of ground truth data, which is difficult to obtain for a real world setting. In order to overcome this dependency, we instead propose the application of self-supervised training, which is able to leverage the vast amount of unlabelled real world data.

2.2. Self-Supervised Optical Flow

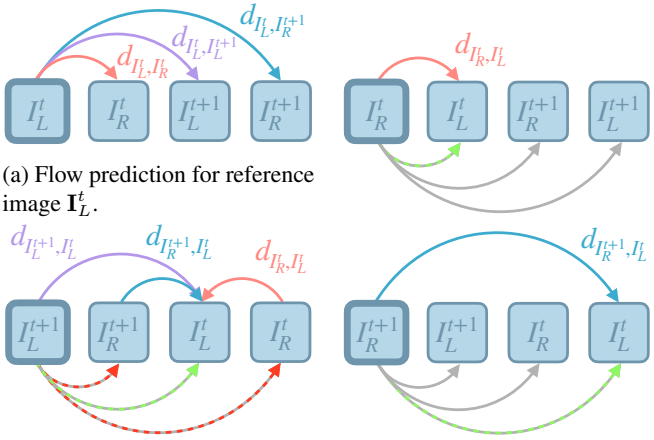
Self-supervised methods for optical flow prediction are mostly based on the photometric similarity of warped frames [10, 11, 12]. However, one obstacle in the usage of the photometric loss is its lack of invariance towards monotonic illumination changes, which appear commonly in real world scenarios. The authors of [2] approach this issue by utilizing a ternary census loss. They further leverage a bidirectional flow setup for occlusion detection. Moreover, approaches like [13, 14] apply geometric constraints or reason about the rigidity of objects in the scene. General improvements for self-supervision in optical flow settings are described in [3] and involve among other things cost volume normalization addressing the problem of vanishing feature activation. Our approach as well relies on the photometric loss and best practices in self-supervised optical flow prediction, but further exploits the availability of stereo image sequences to improve these concepts in the context of scene flow prediction.

2.3. Self-Supervised Scene Flow

The current research regarding scene flow prediction exhibits a large lack in self-supervised methods. The work of [15] focuses on monocular scene flow prediction based on a photometric and 3D point reconstruction loss. However, the monocular input leads to a weakly posed problem, compared to the stereo setting. Furthermore, [16] and [17] discuss self-supervision for point clouds. However, the sensing of point clouds is expensive and often results in an insufficient density. Due to these reasons, we solve – for the first time – the problem of self-supervised scene flow prediction in a stereo camera setting.

3. SELF-SUPERVISED SCENE FLOW PREDICTION

We assume a standard stereo setting for scene flow prediction, in which four images $\mathbf{I}_L^t, \mathbf{I}_R^t, \mathbf{I}_L^{t+1}, \mathbf{I}_R^{t+1}$ (left and right frames at two time steps) are consumed to produce a four dimensional scene flow prediction $\mathbf{s} = (u, v, disp_t, disp_{t+1})^T$ consisting of the optical flow field and two disparity maps [5] (c.f. Figure 1). By this, our self-supervised framework enables the training of a broad range of architectures for scene flow prediction on unlabelled data. One deep neural network in compliance with this definition is PWOC-3D [9], which we choose as our base architecture, because of its simplicity and fast run time. Additionally, our method leverages the stereo setup to compute bidirectional displacements for the estimation of occlusions. This way, the self-supervised loss can be easily applied symmetrically for multiple reference images.



(a) Flow prediction for reference image $\mathbf{I}_L^t$.

(b) Flow prediction for reference image $\mathbf{I}_L^{t+1}$ and required inverse displacements for (a). The predictions in red can not be used as inverse displacements directly.

(c) Flow predictions for reference images with spatial *flipping*, which add the missing inverse displacements in (b).

Fig. 1: Visualization of the backward and forward displacement calculation with and without *flipping*. Corresponding forward and backward displacement are colored in the same way. $d_{\mathbf{I}_1,\mathbf{I}_2}$ describes the displacement from image $\mathbf{I}_1$ towards $\mathbf{I}_2$.

3.1. Self-Supervised Loss

Our self-supervised loss is based on the photometric reconstruction loss, i.e. the difference between the reference and warped image based on the scene flow prediction. According to [2], this loss can be formulated, for two images $\mathbf{I}_1, \mathbf{I}_2$ which are related by the two dimensional displacement $d_{\mathbf{I}_1,\mathbf{I}_2}$, over all pixel positions $\mathbf{x}$ in the following way:

$$L(\mathbf{I}_1, \mathbf{I}_2) = \sum_{\mathbf{x}} \Big(\mathbf{O}(\mathbf{x}) \cdot \rho\Big(\Phi\big(\mathbf{I}_1(\mathbf{x}), \mathbf{I}_2(\mathbf{x} + d_{\mathbf{I}_1,\mathbf{I}_2}(\mathbf{x}))\big)\Big) + \lambda\big(1 - \mathbf{O}(\mathbf{x})\big)\Big), \tag{1}$$

with $\rho(x)$ as the Charbonnier function [18] and $\Phi(\mathbf{I}_1(\mathbf{x}), \mathbf{I}_2(\mathbf{x}'))$ as the photometric difference at the corresponding pixel $\mathbf{x}$ and $\mathbf{x}'$. In order to introduce invariance for monotonic illumination changes, we apply the ternary census transform [19] on the images and compute

the Hamming Distance. Furthermore, we utilize a binary occlusion mask $\mathbf{O}$, in order to mask occluded pixels. We determine this occlusion mask by applying a consistency check [20] (see Section 3.2). To avoid the trivial solution of occluding everything, we add a constant penalty $\lambda = 12.4$ for all occluded pixels [2]. In addition, we use a second-order smoothness constraint [21]. To improve the self-supervision we also apply cost volume normalization [3] on the corresponding layer in our network as a means to support improved feature activation. The reconstruction loss over two images is afterwards computed for all image pairs between the reference image $\mathbf{I}_L^t$ and the rest of the input images:

$$L_{pairs}(\mathbf{I}_L^t) = L(\mathbf{I}_L^t, \mathbf{I}_R^t) + L(\mathbf{I}_L^t, \mathbf{I}_L^{t+1}) + L(\mathbf{I}_L^t, \mathbf{I}_R^{t+1}). \quad (2)$$

3.2. Bidirectional Displacements

The forward displacements from the reference image $\mathbf{I}_L^t$ towards all the other images are the network's prediction for the default input. A bidirectional consistency check for the estimation of occlusions requires the computation of the inverse displacements, based on an altered input. The corresponding standard approach in optical flow is using an alteration of the time-wise order of the input images. In our case, this results in the following input for the network $\mathbf{I}_L^{t+1}$, $\mathbf{I}_R^{t+1}$, $\mathbf{I}_L^t$, and $\mathbf{I}_R^t$, in which $\mathbf{I}_L^{t+1}$ is the reference image. To derive corresponding backward displacements for the forward scene flow prediction, additional warping operations become necessary (c.f. Figure 1). These not only require a greater computational overhead but also introduce holes in the predictions as well as an additional interpolation error. In order to avoid this obstacle, we propose spatial *flipping* of the input, i.e. not only altering the temporal but also the spatial order (left-right). This allows us to predict scene flows for all images and therefore directly retrieve the corresponding inverse displacements from a suitable flow (see Figure 1c). Based on the predicted scene flows our approach is moreover easily extendable to consider losses for additional reference images, compared to supervised approaches, which only focus on the reference image $\mathbf{I}_L^t$. Favoring $\mathbf{I}_L^t$ as reference view has no conceptual reasoning, since optical challenges like occlusion, specular reflections, noise etc. can occur in all frames [22]. Furthermore, multiple reference images provide more information and thus enforce further restrictions on the correctness of the scene flow prediction. If multiple reference images I_{ref} are considered, the overall loss consists of the sum of pairwise losses for all of them:

$$L_{total} = \sum_{\mathbf{I} \in I_{ref}} L_{pairs}(\mathbf{I}). \quad (3)$$

Table 1: Ablation study for different components of our photometric loss on our validation set. We assess the influence of cost volume normalization, the choice of reference images, as well as the application of *flipping*.

Costvol. Norm.	Ref. Image(s)	Flipping	EPE	KOE
✓	I_L^t	✗	14.26	54.14
✗	I_L^t, I_R^t	✗	19.78	47.06
✓	I_L^t, I_R^t	✗	**10.79**	36.19
✓	I_L^t, I_R^t	✓	12.22	**30.35**
✓	$I_L^t, I_R^t, I_L^{t+1}, I_R^{t+1}$	✓	11.93	31.32

4. EXPERIMENTS AND RESULTS

4.1. Details

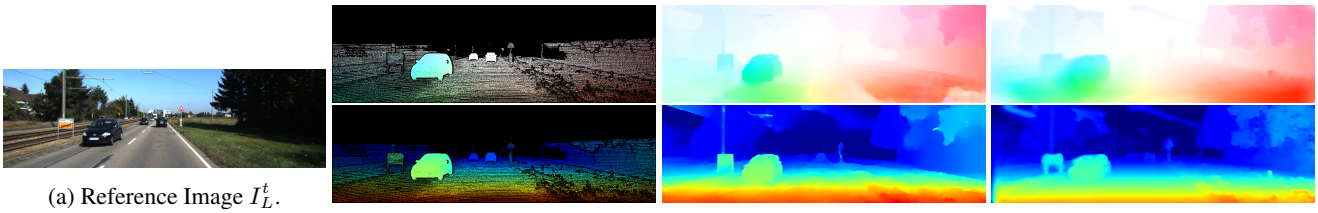
Since our work focuses primarily on real world scenarios, the KITTI raw data set [4] is a natural choice for a training data set. Furthermore, we utilize the annotated KITTI 2015 data set [5] for validation and fine-tuning using the same splitting as in [9]. In order to ensure unbiased and comparable results, we remove sequences from the training data, that also appear in the validation data, resulting in 35666 training samples. In the fashion of [9], we further use the Adam [23] optimizer and apply a learning rate of 0.00017 with a batch size of $b = 3$. For the loss we utilize the recommended parameter settings in [2]. For all our experiments, we evaluate the average endpoint error (EPE [px]) and the KITTI outlier error (KOE [%]) [5] of the predicted scene flow field.

4.2. Ablation Study

As evident from Table 1, additional reference images enhance the performance of the model due to the improved representation of the relationship between the predicted scene flow and the corresponding image pairs which enforces further constraints on the correctness of the prediction. However, if *flipping* is applied its effect decreases, leading to similar results between 2 and 4 reference images due to the great amount of information already included in the loss. Furthermore, the results show, that the utilization of *flipping* drastically improves the performance. This is caused by the supplementary information it offers during the training process which enhances the networks understanding of the relationship between bidirectional flows. We can also observe a clearly positive effect from applying cost volume normalization.

4.3. Comparison to Supervised Pre-Training

Table 2 shows that the photometric approach is able to drastically outperform the supervised pre-training on synthetic data with almost 20 percentage points less outliers and about 2px lower EPE. This indicates the strong influence of the domain gap between synthetic and real images. The KITTI raw data



(a) Reference Image I_L^t.

(b) Ground Truth (visually enhanced).

(c) Supervised pre-training on synthetic data.

(d) Pretrained with our Self-SuperFlow.

Fig. 2: Visualization of optical flow and first disparity for the supervised pre-training and our photometric loss on a KITTI validation sample before fine-tuning.

Table 2: Comparison between the supervised training method (as in PWOC-3D [9]) and our self-supervised method on our validation split, before and after supervised fine-tuning.

Data	Supervised	Stage	Epochs	EPE-all	KOE-all
synthetic	✓	pre	760	14.52	48.91
real	✗	pre	24	**12.22**	**30.35**
real	✓	fine	125	**3.52**	13.76
real	✓	fine	125	3.92	**13.70**

Table 3: Comparison of self-supervised and supervised methods on the KITTI test set (online benchmark).

	Method	D1-all	D2-all	Fl-all	SF-all
stereo	Self-SuperFlow-ft (Ours)	**4.66**	8.65	**12.12**	**14.73**
	PWOC-3D-ft [9]	5.13	**8.46**	12.96	15.69
	Self-SuperFlow (Ours)	8.11	21.57	23.67	28.71
mono	Multi-Mono-SF-ft [24]	22.71	26.51	13.37	33.09
	Self-Mono-SF-ft [15]	22.16	25.24	15.91	33.88
	Multi-Mono-SF [24]	30.78	34.41	19.54	44.04
	Self-Mono-SF [15]	34.02	36.34	23.54	49.54

set [4] used for our self-supervised method resembles the target domain much closer and is further able to provide higher versatility and variation compared to the synthetic data used in the supervised pre-training. A more detailed evaluation reveals that our self-supervised approach is inferior in terms of EPE in the foreground areas. The reason for this are occlusions, which lead to a lack of information that can not be overcome by the self-supervised method. Furthermore, foreground areas generally pose a greater challenge since they typically involve higher magnitudes of displacements. After fine-tuning, the self-supervised training is still capable of achieving similar results compared to the supervised one. Figure 2 further shows, that the supervised method produces sharper edges, which may be hindered by the smoothness constraint in the self-supervised method. However it can be seen, that the photometric loss is able to capture more detail especially in background areas. In addition, our self-supervised method requires only 24 epochs ($\sim 0.285M$ iterations with batches of 3) of pre-training, while the supervised approach requires 760 training epochs ($\sim 7.84M$ iterations with batches of 2). This validates the improved convergence capabilities when real world data is used for the training.

4.4. Comparison to SotA on the KITTI Benchmark

Table 3 shows the results of our method compared to PWOC-3D and other self-supervised approaches on the KITTI scene flow benchmark[1]. It is evident, that after fine-tuning our approach achieves a superior overall scene flow outlier rate compared to the other ones. It even outperforms the supervised training of PWOC-3D by 1 %, showing improved generalization capabilities after fine-tuning when real world data is used for the pre-training. Especially the result in the disparity prediction surpasses all the other methods. To no surprise, our method cuts the outlier rates by half compared to the monocular methods Self-Mono-SF [15] and Multi-Mono-SF [24].

[1] http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php

5. CONCLUSION

In this work we approach the lack of real world ground truth data and the inherent representation gap due to the usage of synthetic training images by introducing a novel self-supervised training method for scene flow prediction in stereo sequences, which can be applied to a vast range of recent networks. Our approach is based on a bidirectional reconstruction loss as well as a forward-backward consistency check for occlusion awareness. We further propose an innovative strategy to derive the inverse displacement in a scene flow setting using *flipping* i.e. spatial variations of the input images.

Without fine-tuning, our self-supervised approach outperforms the equivalent supervised training by over 20 percentage points less outliers on the validation set, proving the significant influence of the representation gap on the training performance. We were further able to show a faster convergence during training and an improved generalization capability of our method compared to the supervised one.

6. REFERENCES

[1] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox, "FlowNet: Learning Optical Flow with Convolutional Networks," in *International Conference on Computer Vision (ICCV)*, 2015.

[2] Simon Meister, Junhwa Hur, and Stefan Roth, "UnFlow: Unsupervised Learning of Optical Flow with a Bidirectional Census Loss," in *Conference on Artificial Intelligence (AAAI)*, 2018.

[3] Rico Jonschkowski, Austin Stone, Jonathan T Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova, "What matters in unsupervised optical flow," in *European Conference of Computer Vision (ECCV)*, 2020.

[4] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets Robotics: The KITTI Dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[5] Moritz Menze and Andreas Geiger, "Object scene flow for autonomous vehicles," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[6] Evan Herbst, Xiaofeng Ren, and Dieter Fox, "RGB-D Flow: Dense 3-D motion estimation using color and depth," in *International Conference on Robotics and Automation (ICRA)*, 2013.

[7] Tali Basha, Y. Moses, and N. Kiryati, "Multi-view scene flow estimation: A view centered variational approach," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[8] C. Vogel, K. Schindler, and S. Roth, "Piecewise Rigid Scene Flow," in *International Conference on Computer Vision (ICCV)*, 2013.

[9] Rohan Saxena, René Schuster, Oliver Wasenmüller, and Didier Stricker, "PWOC-3D: Deep Occlusion-Aware End-to-End Scene Flow Estimation," in *Intelligent Vehicles Symposium (IV)*, 2019.

[10] Aria Ahmadi and Ioannis Patras, "Unsupervised convolutional neural networks for motion estimation," in *International Conference on Image Processing (ICIP)*, 2016.

[11] J Yu Jason, Adam W Harley, and Konstantinos G Derpanis, "Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness," in *European Conference on Computer Vision (ECCV)*, 2016.

[12] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander G Hauptmann, "Guided optical flow learning," *arXiv preprint arXiv:1702.02295*, 2017.

[13] Stefano Alletto, Davide Abati, Simone Calderara, Rita Cucchiara, and Luca Rigazio, "TransFlow: Unsupervised Motion Flow by Joint Geometric and Pixel-level Estimation," *arXiv preprint arXiv:1706.00322*, 2017.

[14] Zhichao Yin and Jianping Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[15] Junhwa Hur and Stefan Roth, "Self-supervised monocular scene flow estimation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[16] Himangi Mittal, Brian Okorn, and David Held, "Just go with the flow: Self-supervised scene flow estimation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[17] Victor Zuanazzi, Joris van Vugt, Olaf Booij, and Pascal Mettes, "Adversarial self-supervised scene flow estimation," in *International Conference on 3D Vision (3DV)*, 2020.

[18] Andrés Bruhn, Joachim Weickert, and Christoph Schnörr, "Lucas/Kanade Meets Horn/Schunck: Combining Local and Global Optic Flow Methods," *International Journal of Computer Vision (IJCV)*, 2005.

[19] Ramin Zabih and John Woodfill, "Non-parametric local transforms for computing visual correspondence," in *European Conference on Computer Vision (ECCV)*, 1994.

[20] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer, "Dense point trajectories by gpu-accelerated large displacement optical flow," in *European Conference of Computer Vision (ECCV)*, 2010.

[21] Werner Trobin, Thomas Pock, Daniel Cremers, and Horst Bischof, "An unbiased second-order prior for high-accuracy motion estimation," in *Pattern Recognition*, 2008.

[22] Christoph Vogel and Stefan Roth, "3D Scene Flow Estimation with a Piecewise Rigid Scene Model," *International Journal of Computer Vision (IJCV)*, 2015.

[23] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[24] Junhwa Hur and Stefan Roth, "Self-supervised multi-frame monocular scene flow," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.