Unsupervised Learning of Multi-Frame Optical Flow with Occlusions

Joel Janai[1,3]([✉]), Fatma Güney[4], Anurag Ranjan[2], Michael Black[2], and Andreas Geiger[1,3]

[1] Autonomous Vision Group, MPI for Intelligent Systems, Tübingen, Germany
joel.janai@tue.mpg.de
[2] Perceiving Systems Department, MPI for Intelligent Systems, Tübingen, Germany
[3] University of Tübingen, Tübingen, Germany
[4] Visual Geometry Group, University of Oxford, Oxford, England

Abstract. Learning optical flow with neural networks is hampered by the need for obtaining training data with associated ground truth. Unsupervised learning is a promising direction, yet the performance of current unsupervised methods is still limited. In particular, the lack of proper occlusion handling in commonly used data terms constitutes a major source of error. While most optical flow methods process pairs of consecutive frames, more advanced occlusion reasoning can be realized when considering multiple frames. In this paper, we propose a framework for unsupervised learning of optical flow and occlusions over multiple frames. More specifically, we exploit the minimal configuration of three frames to strengthen the photometric loss and explicitly reason about occlusions. We demonstrate that our multi-frame, occlusion-sensitive formulation outperforms existing unsupervised two-frame methods and even produces results on par with some fully supervised methods.

1 Introduction

Accurate estimation of optical flow is a long standing goal in computer vision, yet certain aspects of the problem remain largely unsolved to date. This can be attributed to the large degree of ambiguities inherent to this ill-posed problem which can only be resolved using prior knowledge about the appearance and motion of image sequences.

Early approaches addressing the optical flow problem [1,2] integrate simple local smoothness assumptions about the optical flow field using variational optimization. To overcome the limitations of local priors, patch-based MRF for-

---

mulations [3–5] and semantics [6,7] have been exploited. More recently, deep neural networks [8–11] have been successfully applied to the optical flow problem. Learning to solve optical flow in an end-to-end fashion from examples is attractive as deep neural networks allow for learning more complex hierarchical flow representations directly from annotated data. However, training such models requires large datasets and obtaining ground truth for real images is challenging as labeling dense correspondences by hand is intractable. Unlike stereo where active sensors such as structured light or laser scanners can be used, no other technology is able to directly deliver optical flow ground truth [12]. Thus, existing approaches train primarily on synthetic data [8,13,14]. However, creating data from a distribution that resembles natural scenes is a hard problem on its own.
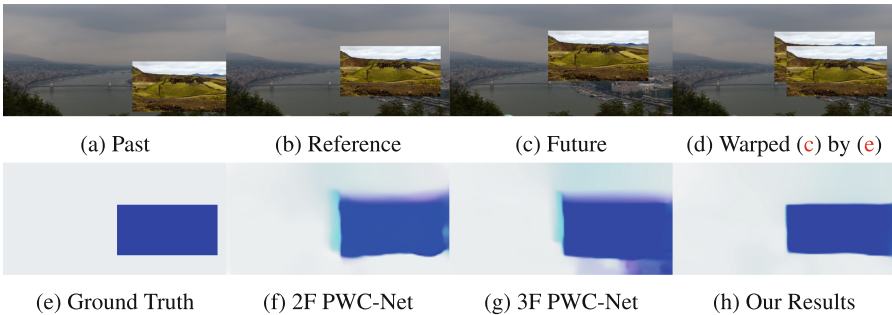

(a) Past   (b) Reference   (c) Future   (d) Warped (c) by (e)

(e) Ground Truth   (f) 2F PWC-Net   (g) 3F PWC-Net   (h) Our Results

Fig. 1. Motivation. Unsupervised optical flow estimation is challenging as commonly used photometric terms are violated in occluded regions. This example from our RoamingImages dataset illustrates the problem of *ghosting effects* (d) when warping the target frame (c) according to the true flow (e). Classical two-frame approaches produce blurry results near occlusion boundaries (f). Using multiple frames without occlusion reasoning neither alleviates the problem (g). In contrast, our multi-frame model with explicit occlusion reasoning leads to accurate flow estimates with sharp boundaries (h).

Alternatively, optical flow can be treated as an unsupervised learning problem. In the unsupervised case, a photometric loss is minimized [15–20], measuring how well the predicted flow warps the target image to the reference frame. Particularly problematic in this setting are occluded regions [19,20] which provide misleading information to the photometric loss function. This problem is illustrated in Fig. 1 with an example from the synthetic 'RoamingImages' dataset which we have created based on randomly moving image patches from Flickr. The photometric loss compares the reference image (Fig. 1(b)) to the target image that is warped according to the optical flow estimate (Fig. 1(d)). Note that occluded regions in the target image cannot be recovered correctly even when using the ground truth optical flow field (Fig. 1(e)). Instead, the so-called *ghosting effects* occur, i.e., parts of the occluder remain visible in the occluded regions. Recent works [19,20] propose to exclude these regions in the photo-

metric loss by inferring occluded regions using the backward flow, i.e., the flow from the target frame to the reference frame. However, these approaches depend heavily on an accurate flow prediction and use heuristics (e.g., thresholding) to infer occlusions.

We propose to model temporal relationships over multiple frames in order to learn optical flow and occlusions jointly. For this purpose, we extend the two-frame architecture proposed in [11] to multiple frames. We estimate optical flow in both past and future direction together with an occlusion map within a temporal window of three frames. Our unsupervised loss evaluates the warped images from the past and the future based on the estimated flow fields and occlusion map. In addition to typical spatial smoothness constraints, we introduce a constant velocity constraint within the temporal window. This allows to reason about occlusions in a principled manner while leveraging temporal information for more accurate optical flow prediction in occluded regions.

We perform ablation studies on our RoamingImages dataset considering two-frame and multi-frame formulations without occlusion modeling as our baselines. In addition, we evaluate our approach on KITTI 2015 [21,22] and MPI Sintel [14]. Surprisingly, our model trained only on the simplistic RoamingImages dataset outperforms all existing unsupervised optical flow approaches trained on FlyingChairs [15,18,19]. By unsupervised fine-tuning on the respective training sets, we further improve our results, reducing the gap to several supervised methods. We will make the code and the trained models available upon the publication.

We summarize the contributions of this paper as follows:

- We propose a novel **unsupervised, multi-frame optical flow** formulation which estimates past and future flow within a three-frame temporal window.
- By explicitly reasoning about **occlusions**, we increase the fidelity of the photometric loss, resulting in sharper boundaries (Fig. 1(h)) in comparison to two-frame (Fig. 1(f)) as well as three-frame formulations without occlusions (Fig. 1(g)).
- We demonstrate that **temporal constraints** enable more accurate optical flow predictions in occluded regions compared to just spatial propagation, as in all existing unsupervised two-frame optical flow formulations.

2   Related Work

Classic Multi-Frame Optical Flow: While the majority of optical flow methods use two input frames, few works have exploited the properties of temporal coherence in video sequences. Early approaches to multi-frame optical flow use phase-based representations for encoding the local image structure [23,24]. Later, variational optical flow approaches [1,2] have been extended to multiple frames via spatio-temporal regularizers [12,25–29] using either a constant velocity prior [12,30–33] or assuming constant acceleration [34,35]. In addition to temporal constraints, multi-frame formulations also allow to reason about the visibility of a pixel. Occluded regions are particularly problematic in unsupervised learning

of optical flow due to the weak photometric terms used for training. However, to the best of our knowledge, neither temporal constraints nor occlusion reasoning with multiple frames have been explored in the context of unsupervised learning. This paper presents the first approach to leverage a multi-frame formulation for learning optical flow and occlusions in an unsupervised fashion. More specifically, we focus on the minimal case of three frames which allows us to reason about the visibility of a pixel while expecting only little appearance changes that mostly adhere to the brightness constancy assumption.

Deep Neural Networks: In recent years, end-to-end approaches for learning the optical flow problem have shown great promise [8–11]. Typically, a model composed of encoder and decoder modules takes two stacked consecutive frames as input. This kind of architecture for optical flow was first proposed in FlowNet [8] and extended in FlowNet2 [10] by stacking multiple encoder-decoder networks one after the other. Following the coarse-to-fine idea in traditional optical flow estimation, Ranjan et al. [9] (SPyNet) use warped images over multiple scales for handling large displacements. Sun et al. [11] (PWC-Net) combine different ideas from optical flow and stereo matching by training a shallow Siamese network and constructing a cost volume at different scales.

In this paper, we build on PWC-Net since their framework is lightweight, produces state-of-the-art results and allows for an elegant integration of our multi-frame formulation. In addition to optical flow, our model also reasons about occlusions. In contrast to the fully supervised setting [8–11], we train our model without ground truth flow.

Unsupervised Learning: The dependency of deep neural networks on large annotated datasets has recently motivated the development of unsupervised learning techniques. Impressive results have been demonstrated for single image depth prediction [16,36–39], ego-motion estimation [16,39,40] and optical flow [15–20,41,42]. In a typical unsupervised optical flow framework, a photometric loss is used in combination with a smoothness loss for untextured regions [15–20,41,42]. More specifically, the target image is warped according to the predicted flow and compared to the reference image using a photometric loss. Typically, an encoder-decoder network [15,17–20] is used. Pătrăucean et al. [17] combine the simple encoder-decoder network with a convolutional LSTM to incorporate information from previous frames. We also use a photometric loss over multiple frames but instead of using an LSTM, we modify the network architecture proposed in [11] to directly encode the temporal relationship with a constant velocity assumption over three frames.

Recently, [19,20] proposed to exclude occluded regions from the photometric loss to avoid misleading information. While both of them jointly learn the forward and backward flow, Meister et al. [20] use a forward-backward consistency check and Wang et al. [19] create a range map with the backward flow, counting the correspondences for each pixel in the reference frame. However, both approaches use a heuristic to obtain the final occlusion map. Instead of using a heuristic, we estimate the occlusion maps jointly with the optical flow. We relate flow and occlusion estimates in our photometric loss by weighting information

from the future and the past according to occlusion estimates. This joint formulation allows us to train our occlusion-aware model from scratch in contrast to [20] that requires pre-training without occlusion reasoning. Another recent work on unsupervised learning of depth and ego-motion [39] predicts explainability masks to exclude dynamic objects and occlusions using a photometric loss function. While [39] only addresses static scenes, we target the general unconstrained optical flow problem and learn to jointly predict flow and occluded regions in this setting.

3   Method

In this paper, we propose an approach for unsupervised learning of optical flow and occlusions by leveraging multiple frames. In unsupervised learning of optical flow, only the photometric loss provides guidance. The photometric loss warps the target frame according to the flow estimate and compares the warped target frame to the reference frame. Local ambiguities caused by untextured regions are handled with an additional spatial smoothness constraint that propagates information between neighboring pixels. However, learning optical flow in an unsupervised fashion is complicated due to ambiguities caused by non-lambertian reflectance, occlusions, large motions and illumination changes. Considering multiple frames can help to resolve some of the ambiguities, in particular those caused by occlusions. We thus propose a multi-frame formulation to train a convolutional neural network to predict flow fields and occlusions jointly.

3.1   Notation

We first introduce our notation. Let $\mathcal{I} = \{\mathbf{I}_P, \mathbf{I}_R, \mathbf{I}_F\}$ denote three consecutive RGB frames $\mathbf{I}_t \in \mathbb{R}^{W \times H \times 3}$. Our goal is to predict the optical flow $\mathbf{U}_F \in \mathbb{R}^{W \times H \times 2}$ from reference frame $\mathbf{I}_R$ to future frame $\mathbf{I}_F$ while leveraging the past frame $\mathbf{I}_P$. In this short temporal window, we assume the motion to be approximately linear. The simplest way to enforce a linear motion is using a hard constraint by predicting only one flow field and warping both images $\mathbf{I}_P, \mathbf{I}_F$ to the reference image $\mathbf{I}_R$ according to this flow field for computing the photometric loss. However, realistic scenes usually contain more complex motions which violate this hard constraint (e.g., road surface in KITTI). Therefore, we formulate a soft constraint by predicting two optical flow fields and encouraging constant velocity: We denote $\mathbf{U}_F$ the flow field from reference frame $\mathbf{I}_R$ to future frame $\mathbf{I}_F$, and $\mathbf{U}_P \in \mathbb{R}^{W \times H \times 2}$ the flow field from reference frame $\mathbf{I}_R$ to past frame $\mathbf{I}_P$.

Regardless of the motion model, photo-consistency is violated in occluded regions. Considering three frames allows to resolve this problem by reasoning about occlusions in a data-driven fashion. Let us consider a pixel $\mathbf{p}$ in reference frame $\mathbf{I}_R$. Note that by definition the pixel is visible in the reference frame. Thus, there are only three possible cases: Either it is visible in all frames, or it has been occluded in the past, or it becomes occluded in the future. While there exists a possible fourth state, i.e., when a pixel is solely visible in the

reference frame, this is a very unusual case that rarely occurs in practice and therefore can be discarded. Thus, the occlusion of each pixel can be represented with three states and we can always exploit information by either considering the future or the past. More formally, we model occlusions by introducing a continuous occlusion variable $\mathbf{O} \in [0,1]^{W \times H \times 2}$ at every pixel which allows to correctly evaluate the photometric loss by reducing the importance of occluded pixels. Let $\mathbf{O(p)} \in [0,1]^2$ denote the occlusion at pixel $\mathbf{p}$ where $\|\mathbf{O(p)}\|_1 = 1$. If $\mathbf{O(p)} = (1,0)$, we consider $\mathbf{p}$ as backward occluded (i.e., occluded in the previous frame), if $\mathbf{O(p)} = (0,1)$, pixel $\mathbf{p}$ is forward occluded and if $\mathbf{O(p)} = (0.5, 0.5)$, pixel $\mathbf{p}$ is visible in all frames.

We propose to estimate $\mathbf{U}_F$, $\mathbf{U}_P$ and $\mathbf{O}$ jointly using a neural network and enforcing $\|\mathbf{O(p)}\|_1 = 1$ with a softmax at the last layer of the network.
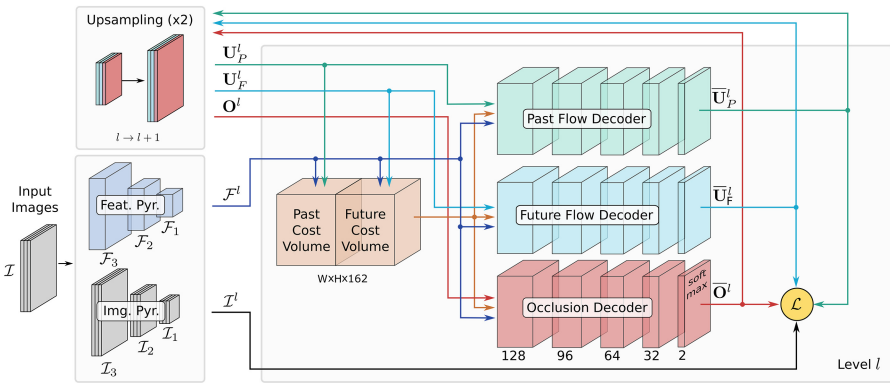

Fig. 2. Network architecture. Given the input sequence $\mathcal{I}$, we construct an image and a feature pyramid. The optical flow is estimated in a coarse-to-fine manner: at level l, two cost volumes are constructed from the features $\mathcal{F}^l$ of the past and future frame, warped according to the current optical flow estimates $\mathbf{U}_P^l$ and $\mathbf{U}_F^l$, respectively. The two cost volumes are decoded resulting in the past flow $\overline{\mathbf{U}}_P^l$, future flow $\overline{\mathbf{U}}_F^l$ and an occlusion map $\overline{\mathbf{O}}^l$ at level l. The estimations are passed to the upsampling block to yield inputs for the next level $l+1$ of the pyramid. See text for details.

3.2 Network Architecture

The recently proposed PWC-Net architecture [11] borrows ideas from the stereo literature and constructs a cost volume from the features of the reference frame and the warped features of the future frame. Finally, a fully convolutional decoder returns the optical flow for each level that is used to warp the features to the next level. This results in a compact and discriminative representation producing state-of-the-art performance.

Inspired by the supervised two-frame PWC-Net model, we develop our unsupervised multi-frame and occlusion aware formulation illustrated in Fig. 2. Similar to PWC-Net, we estimate the flow fields and occlusion maps in a coarse-to-fine manner. The first modification we make is to add the past frame to the

image and feature pyramids. In the original PWC-Net, a cost volume is constructed based on the features of the reference frame and the features of the target frame warped according to the flow estimate. In contrast, we construct two cost volumes: one for the past and one for the future frame. The two separate cost volumes allow our network to detect occlusions and choose the relevant information for accurate optical flow estimation. Finally, we use three separate decoders for future flow, past flow and occlusion map, respectively. The cost volumes are stacked together and form the input to the decoders. We upsample past flow, future flow and occlusion map predictions from the previous level and provide them accordingly as input to the decoders together with the cost volume and the features of the reference frame. For all three decoders, we use the decoder architecture proposed in [11], just for the occlusion decoder we add a softmax at the end.

Our architecture with two flow decoders is designed to encourage constant velocity as a soft constraint. We also experiment with an architecture using one flow decoder for both directions. In that case, the inverse future flow is treated as the estimation for past flow. This corresponds to a hard constraint which is useful in cases where the linear assumption always holds, e.g. on our RoamingImages dataset.

3.3   Loss Functions

Our goal is to learn accurate optical flow and occlusions within a temporal window in an unsupervised manner. Let $\boldsymbol{\theta}$ denote the parameters of a neural network which predicts $\mathbf{U}_F(\boldsymbol{\theta})$, $\mathbf{U}_P(\boldsymbol{\theta})$ and $\mathbf{O}(\boldsymbol{\theta})$ from the input images $\mathcal{I}$. Our loss $\mathcal{L}(\boldsymbol{\theta})$ is a linear combination of a photometric loss $\mathcal{L}_P(\boldsymbol{\theta})$, smoothness constraints $\mathcal{L}_{S_P}(\boldsymbol{\theta}), \mathcal{L}_{S_F}(\boldsymbol{\theta}), \mathcal{L}_{S_O}(\boldsymbol{\theta})$, a constant velocity constraint $\mathcal{L}_{CV}(\boldsymbol{\theta})$ and an occlusion prior $\mathcal{L}_O(\boldsymbol{\theta})$:

$$\mathcal{L} = \mathcal{L}_P + \mathcal{L}_{S_F} + \mathcal{L}_{S_P} + \mathcal{L}_{S_O} + \mathcal{L}_{CV} + \mathcal{L}_O \tag{1}$$

For clarity, we dropped the dependency on the parameters $\boldsymbol{\theta}$ and the relative weights of the loss functions. While the first two terms have been frequently employed by unsupervised methods before [15–20,41,42], we extend this formulation to the multi-frame scenario with a simple but effective linear motion model and proper handling of occlusions. In the following, we describe each individual term in detail.

Photometry: In unsupervised optical flow estimation, supervision is achieved by warping the images according to the predicted optical flow and comparing the intensity or color residuals. Unlike existing approaches [15–20], we take advantage of multiple frames to strengthen the photometric constraint. Similar to [19,20], our model takes occlusions into account. While these methods use simple heuristics based on thresholding to obtain occlusion maps for masking, we directly model occlusions in our formulation and use them to weight the contribution of future and past estimates. Our approach is able learn more

sophisticated models which allow for more accurate occlusion reasoning. Moreover, our approach allows the network to avoid errors in occluded regions since a pixel is by definition always visible in at least two frames. More formally, we formulate our photometric loss as

$$\mathcal{L}_P = \sum_{\mathbf{p}\in\Omega} \mathbf{O}^{(2)}(\mathbf{p}) \cdot \delta\left(\hat{\mathbf{I}}_P\left(\mathbf{p} + \mathbf{u}_P\left(\mathbf{p}\right)\right), \mathbf{I}_R\left(\mathbf{p}\right)\right) \tag{2}$$
$$+ \sum_{\mathbf{p}\in\Omega} \mathbf{O}^{(1)}(\mathbf{p}) \cdot \delta\left(\hat{\mathbf{I}}_F\left(\mathbf{p} + \mathbf{u}_F\left(\mathbf{p}\right)\right), \mathbf{I}_R\left(\mathbf{p}\right)\right)$$

where Ω denotes the domain of the reference image $\mathbf{I}_R$, $\mathbf{u}_P$ and $\mathbf{u}_F$ denote the past and future flow at pixel $\mathbf{p}$, and $\mathbf{O}^{(i)}(\mathbf{p})$ denotes the i'th component of occlusion variable $\mathbf{O}(\mathbf{p})$. Instead of handling occlusions in the warping function, we instead use bilinear interpolation for warping [43] and a robust function $\delta(\cdot, \cdot)$, detailed below, to measure the photometric error between the warped images $\hat{\mathbf{I}}_{P/F}$ and the reference image $\mathbf{I}_R$. Afterwards, we use our occlusion estimates to weight the photometric errors accordingly. If a pixel $\mathbf{p}$ is more likely to be forward occluded, $\mathbf{O}^{(1)}(\mathbf{p}) < \mathbf{O}^{(2)}(\mathbf{p})$, the information from past frame has a larger contribution. Similarly, if a pixel $\mathbf{p}$ is likely backward occluded, $\mathbf{O}^{(1)}(\mathbf{p}) > \mathbf{O}^{(2)}(\mathbf{p})$, the future frame is weighted higher. In the case of pixel $\mathbf{p}$ being visible within the whole window, $\mathbf{O}^{(1)}(\mathbf{p}) \approx \mathbf{O}^{(2)}(\mathbf{p})$, both future and past frames contribute equally. This soft weighting of the data terms ensures that our photometric loss is fully differentiable.

Several photometric error functions have been proposed in the classical optical flow literature. The most popular is the brightness constancy assumption [1] which measures the difference between pixel intensities or colors (Eq. (3)). Instead of the original quadratic penalty function, we use the generalized Charbonnier penalty ρ [44] for robustness against outliers Eq. (5). In realistic scenes with illumination changes, the brightness constancy assumption is often violated and instead a gradient constancy assumption is considered by comparing the gradients of the pixel intensities (Eq. (4)). In this work, we use the brightness constancy assumption when training on synthetic data and the gradient constancy assumption when training on KITTI.

$$\delta_{BC}(I_1, I_2) = \rho\left(I_1 - I_2\right) \tag{3}$$

$$\delta_{GC}(I_1, I_2) = \rho\left(\frac{\partial I_1}{\partial x} - \frac{\partial I_2}{\partial x}\right) + \rho\left(\frac{\partial I_1}{\partial y} - \frac{\partial I_2}{\partial y}\right) \tag{4}$$

$$\rho(\mathbf{x}) = \sum_i \sqrt{x_i^2 + 0.001^2} \tag{5}$$

Smoothness: It is well known that the photometric loss alone does not sufficiently constrain the problem due to the aperture problem and the ambiguity of local appearance. Thus, we add an additional regularizer which encourages smooth flow fields. In particular, we use the following edge aware smoothness loss for $\mathbf{U}_P$:

$$\mathcal{L}_{S_P} = \sum_{\mathbf{p}\in\Omega} \xi\left(\nabla_x \mathbf{I}_R(\mathbf{p})\right) \rho\left(\nabla_x \mathbf{U}_P(\mathbf{p})\right) + \sum_{\mathbf{p}\in\Omega} \xi\left(\nabla_y \mathbf{I}_R(\mathbf{p})\right) \rho\left(\nabla_y \mathbf{U}_P(\mathbf{p})\right) \tag{6}$$

where $\xi(x) = \exp(-\|x\|_2)$ is a contrast sensitive weight to reduce the effect of the smoothness prior at image boundaries, $\nabla_x \mathbf{I}(x,y) = \mathbf{I}(x,y) - \mathbf{I}(x-1,y)$ and $\nabla_x \mathbf{U}$, accordingly, are the backward difference of the image and flow field in spatial direction x. Following [19,20], we can replace the first order smoothness (6) by a second order smoothness which allows piecewise affine flow fields when training on KITTI [45]:

$$\mathcal{L}_{S_P} = \sum_{\mathbf{p} \in \Omega} \xi\left(\nabla_x \mathbf{I}_R(\mathbf{p})\right) \xi\left(\Delta_x \mathbf{I}_R(\mathbf{p})\right) \rho\left(\nabla_x \mathbf{U}_P(\mathbf{p}) - \Delta_x \mathbf{U}_P(\mathbf{p})\right) \qquad (7)$$
$$+ \sum_{\mathbf{p} \in \Omega} \xi\left(\nabla_y \mathbf{I}_R(\mathbf{p})\right) \xi\left(\Delta_y \mathbf{I}_R(\mathbf{p})\right) \rho\left(\nabla_y \mathbf{U}_P(\mathbf{p}) - \Delta_y \mathbf{U}_P(\mathbf{p})\right),$$

Here, $\Delta_x \mathbf{I}(x,y) = \mathbf{I}(x+1,y) - \mathbf{I}(x,y)$ and $\Delta_x \mathbf{U}$, accordingly, denote the forward differences in direction x. The smoothness for the future flow $\mathcal{L}_{S_F}$ is defined accordingly.

Additionally, we introduce a regularizer which encourages similar occlusion states at neighboring pixels:

$$\mathcal{L}_{S_O} = \sum_{\mathbf{p} \in \Omega} \xi\left(\nabla_x \mathbf{I}_R(\mathbf{p})\right) \|\nabla_x \mathbf{O}(\mathbf{p})\|^2 + \sum_{\mathbf{p} \in \Omega} \xi\left(\nabla_y \mathbf{I}_R(\mathbf{p})\right) \|\nabla_y \mathbf{O}(\mathbf{p})\|^2 \qquad (8)$$

Instead of a robust function, we use the squared difference for a stronger penalization of changes between occlusion states.

Constant Velocity: The photometric term and the smoothness term treat the future and past flow separately. In the multi-frame setup, we can go one step further and assume a linear motion model which corresponds to pixels moving with constant velocity within the short temporal window. Despite its simplicity, constant velocity provides a reliable source of information in case of occlusions in addition to spatial smoothness constraints. Under this assumption, the future and past flow should be equal in length but differ in direction. We thus formulate the constant velocity loss as follows:

$$\mathcal{L}_{CV} = \sum_{\mathbf{p} \in \Omega} \rho\left(\mathbf{U}_P(\mathbf{p}) + \mathbf{U}_F(\mathbf{p})\right) \qquad (9)$$

Occlusion Prior: The majority of pixels are typically visible in all frames while occlusions only occur at motion boundaries. We encode this prior as follows:

$$\mathcal{L}_O = -\sum_{\mathbf{p} \in \Omega} \mathbf{O}^{(1)}(\mathbf{p}) \cdot \mathbf{O}^{(2)}(\mathbf{p}) \qquad (10)$$

Note that Eq. (10) is minimized when all pixels are visible (i.e., $\mathbf{O}(\mathbf{p}) = (0.5, 0.5)$).

4   Experimental Results

In this section, we analyze our approach in ablation studies showing the advantages of the multi-frame formulation, occlusion reasoning and constant velocity

assumption. In addition, we compare our method to other unsupervised and supervised methods on established optical flow datasets.

Following the original PWC-Net model [11], we weight the loss function at each level according to the number of pixels, $[0.005, 0.01, 0.02, 0.08, 0.32]$, and scale flow values by 0.05 as in [8,11]. For dataset specific hyper-parameters and settings, please refer to the supplementary. We train our network end-to-end using Adam [46] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We use a batch size of 8 and start with a learning rate of $1e - 4$ for pre-training and $1e - 5$ for fine-tuning. We pre-train our models for 700 K iterations by halving the learning rate after every 200 K iteration. For training, we do not use data augmentation because of the large size of RoamingImages. For evaluation, we consider three standard metrics:

- **End-point Error (EPE)** is defined as the average Euclidean distance between estimated and ground truth flow. We separately report EPE in occluded and visible regions to better analyze the impact of the proposed model components.
- **Average Percentage of Bad Pixels** based on a threshold, i.e. outlier ratio, is used for evaluation on the KITTI 2015 test set.
- **Maximum F-Measure** defined as the weighted harmonic mean of precision and recall for evaluating occlusion estimates.

4.1   Datasets

We use three different datasets in our experiments. We created a simple dataset called 'RoamingImages' to pre-train our model and perform ablation studies. For comparison to other methods, we use two established optical flow datasets in unsupervised setting, the KITTI 2015 dataset [21,22] and MPI Sintel [14].

RoamingImages: Curriculum learning (i.e., pre-training on a simple dataset before fine-tuning on a more complicated one) has proven important when training deep models for optical flow estimation [9,10,15,47]. While deep learning approaches for optical flow typically use the FlyingChairs dataset [8], our multi-frame formulation cannot be trained on this dataset as it provides only two frames per scene. Thus, we have created our own "RoamingImages" dataset by moving a random foreground image in front of a random background image according to a random linear motion as illustrated in Fig. 1. The goal is to gradually learn temporal and occlusion relationships by keeping the geometric relations simple in the beginning. We created 80,000 examples with a resolution of 640×320 that we split into 90% training set and 10% test set.

MPI Sintel: The MPI Sintel dataset [14] was created from the short movie MPI Sintel in Blender and provides ground truth flow and occlusion masks for 1000 image pairs in the training set. Two different rendering passes with different complexity are available ("Clean" and "Final"). In addition, MPI Sintel provides pixel-wise occlusion masks.

KITTI 2015: In contrast to MPI Sintel, the KITTI 2015 dataset [21,22] provides real scenes that were captured from a mobile platform. While the optical flow training set contains only 200 annotated images, the multi-view extension consists of approximately 4000 images. We use all frames except the annotated frames and their neighbors in the training set (frames 9–12) for unsupervised fine-tuning of our model. We will refer to this set as 'KITTI 2015 MV' throughout the remainder of this paper.

Table 1. Ablation Study: We compare our results (Ours) to PWC-Net (Classic) and the multi-frame extension without occlusions (Multi). In addition, we analyze the effect of the constant velocity assumption by turning it off (Ours-None), using a soft constraint (Ours-Soft) or a hard constraint (Ours-Hard). We report flow results using EPE for all (All), non occluded (NOC), and occluded (OCC) pixels.

Method	Frames	Occlusions	Constant velocity	All	NOC	OCC	F-Measure
Classic	2	✗	✗	14.14	9.07	32.03	-
Multi	3	✗	hard	10.11	8.24	18.22	-
Ours-None	3	✓	✗	8.37	6.47	16.26	0.76
Ours-Soft	3	✓	soft	8.17	**6.32**	15.87	0.76
Ours-Hard	3	✓	hard	**6.93**	6.89	**8.55**	**0.83**

4.2   Ablation Study

In this section, we analyze different aspects of our approach on the RoamingImages dataset. More specifically, our goal is to investigate the benefits of our multi-frame formulation with occlusions in comparison to the two-frame case as well as the multi-frame case without occlusion reasoning. In addition, we compare the hard constraint to the soft constraint as well as to the case without any temporal constraints. We list our results in Table 1 and discuss our findings in the next paragraph.

Multi-Frame and Occlusion Reasoning: We first analyze the importance of the multi-frame assumption by training the original two-frame PWC-Net in an unsupervised fashion on RoamingImages (Classic). We then extend PWC-Net to three frames but using only one cost volume without occlusion reasoning (Multi). The multi-frame formulation leads to a significant improvement in the performance reducing the overall EPE from 14.14 to 10.11 (see Table 1). With the multi-frame formulation, even without occlusion reasoning, the error in occluded regions is almost reduced by half. The occlusion reasoning (Ours-Hard) again reduces the error in occluded regions by half compared to the multi-frame formulation without occlusion reasoning (Multi), reaching an overall EPE of 6.93. This clearly shows the benefit of ignoring misleading information in accordance with the occlusion estimates.

Constant Velocity: As explained in Sect. 3, the constant velocity assumption can be enforced in different ways with varying degrees of freedom. In Table 1, we compare the soft constraint case (Ours-Soft) with separate flow fields for future and past optical flow, to the hard constraint case (Ours-Hard) with only one flow estimate for both. In addition, we show results without temporal constraint (Ours-None), i.e., turning off the constant velocity term in the loss while still estimating two flow fields. As evidenced by our results, the hard constraint achieves a significant improvement over the case without temporal constraint on our RoamingImages dataset. In particular, in occluded regions, the error is reduced from 16.26 to 8.55 EPE demonstrating the advantage of the proposed temporal smoothness constraint over a purely spatially regularized model. The soft constraint improves only marginally over the case without temporal constraint demonstrating the benefit of directly encoding the temporal relationship into the model in our restricted scenario.
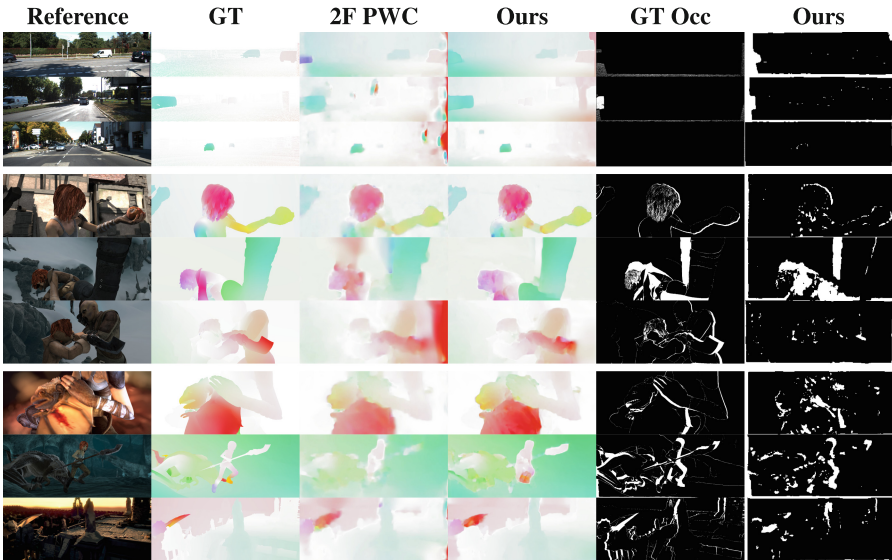

Fig. 3. Qualitative results: We compare our final results (fourth column) to two-frame PWC-Net (third column) on examples from KITTI 2015 (upper three rows) and MPI Sintel Clean (middle three rows) and MPI Sintel Final (bottom three rows). Our model produces better flow estimates with sharper boundaries as well as accurate occlusion estimates (last column).

4.3   Quantitative and Qualitative Results

In Table 2, we compare our method to the state-of-the-art unsupervised approaches DSTFlow [18], UnFlow [20] and OccAwareFlow [19], as well as the leading supervised approaches FlowNet [8], SPyNet [9], FlowNet2 [10], and

Table 2. Quantitative Results: We compare our method to state-of-the-art supervised and unsupervised methods on training and test sets of MPI Sintel and KITTI 2015. We report the EPE except for the KITTI test set where we report the ratio of outliers. Parentheses indicate cases where training was performed on the same dataset and * marks cases where only the annotated samples were excluded from training. Missing entries (–) were not reported for the respective method and bold fonts highlight the best results among supervised and unsupervised methods.

	Methods	MPI sintel clean				MPI sintel final				KITTI 2015			
		Train			Test	Train			Test	Train			Test
		All	NOC	OCC	All	All	NOC	OCC	All	All	NOC	OCC	All
Supervised	FlowNetS [8]	4.50	-	-	7.42	5.45	-	-	8.43	-	-	-	-
	FlowNetS-ft [8]	(3.66)	-	-	6.69	(4.44)	-	-	7.46	-	-	-	-
	SpyNet [9]	4.12	-	-	6.69	5.57	-	-	8.43	-	-	-	-
	SpyNet-ft [9]	(3.17)	-	-	6.64	(4.32)	-	-	8.36	-	-	-	-
	FlowNet2 [10]	2.02	-	-	3.96	3.14	-	-	6.02	10.06	-	-	-
	FlowNet2-ft [10]	**(1.45)**	-	-	4.16	**(2.01)**	-	-	5.74	(2.3)	-	-	11.48%
	PWC-Net [11]	2.55	-	-	-	3.93	-	-	-	10.35	-	-	-
	PWC-Net-ft [11]	(1.70)	-	-	**3.86**	(2.21)	-	-	**5.17**	**(2.16)**	-	-	**9.60%**
Unsupervised	DSTFlow [18]	6.93	5.05	-	10.40	7.82	5.97	-	11.11	24.30	14.23	-	-
	DSTFlow-ft [18]	(6.16)	(4.17)	-	10.41	(6.81)	(4.91)	-	11.27	16.79*	6.96*	-	39%
	UnFlow-CSS [20]	-	-	-	-	7.91	-	-	10.22	8.10*	-	-	23.30%
	OccAwareFlow [19]	5.23	-	-	8.02	6.34	-	-	9.08	21.30	-	-	-
	OccAwareFlow-ft [19]	(4.03)	-	-	7.95	(5.95)	-	-	9.15	8.88*	-	-	31.2%*
	UnFlow-CSS (R) [20]	8.91	-	-	-	10.01	-	-	-	19.26	11.44	-	-
	Ours-Hard	5.38	4.32	11.58	8.35	6.01	4.92	12.42	9.38	15.63	8.80	41.65	48.93%
	Ours-Hard-ft	(6.05)	(4.95)	(12.10)	-	(7.09)	(5.97)	(13.42)	-	11.58*	7.45*	27.29*	-
	Ours-None-ft	(4.74)	(3.60)	(11.42)	-	(5.84)	(4.72)	**(12.66)**	-	6.65*	3.24*	19.33*	-
	Ours-Soft-ft	**(3.89)**	**(2.64)**	**(11.21)**	**7.23**	**(5.52)**	**(4.32)**	(12.87)	**8.81**	**6.59***	**3.22***	**19.11***	**22.94%**

PWC-Net [11] on MPI Sintel and KITTI 2015. In addition, we show qualitative results on KITTI 2015 and MPI Sintel in Fig. 3. We provide an extended version of Table 2 in the supplementary.

While the constant velocity hard constraint works well on the simplistic RoamingImages dataset, more realistic datasets like MPI Sintel and KITTI often exhibit non-linear motions which violate the constant velocity assumption. Therefore, we exploit the soft constraint network on these datasets initialized based on the hard constraint network pre-trained on RoamingImages. More specifically, we copy the parameters of the flow decoder in the pre-trained network to the future and past flow decoders while inverting the sign of the past flow decoder's output. We empirically found this to yield a good initialization for further fine-tuning. Afterwards, we fine-tune our model on the target datasets, i.e.KITTI 2015 MV and MPI Sintel. Note that, during fine-tuning, the model is still trained in an unsupervised fashion. In the following, we present our results in comparison to several state-of-the-art approaches.

Pre-training: Since fine-tuning on a specific dataset makes a big difference, we first consider unsupervised methods without fine-tuning to evaluate our pre-trained model on RoamingImages. Our pre-trained model (Ours-Hard) achieves comparable results on MPI Sintel Clean and significantly outperforms all other unsupervised models without fine-tuning on MPI Sintel Final and KITTI 2015. While the best EPE obtained by a pre-trained unsupervised model is 6.34 on MPI Sintel Final and 21.30 on KITTI 2015, our model achieves an EPE of 6.01 and 15.63, respectively. On MPI Sintel Final, we are even on par with the model of OccAwareFlow fine-tuned on MPI Sintel. This is particularly impressive considering the simplistic dataset used for training our model.

Hard versus Soft Constraint: We compare our hard constraint network to our soft constraint variant to demonstrate the necessity to relax the constant velocity assumption for more complex datasets. While our model with hard constraint (Ours-Hard-ft) improves after fine-tuning on KITTI 2015, its performance is still behind other unsupervised, fine-tuned approaches. On MPI Sintel, the performance decreases after fine-tuning because the constant velocity constraint is wrongly enforced on non-linear motion which frequently occurs in this dataset. Switching to the soft constraint version (Ours-Soft-ft) allows deviations from constant velocity assumption and results in significant improvements on both datasets. For completeness, we include our fine-tuned model without temporal constraint (Ours-None-ft) in the comparison. Similar to Table 1, the performance of the model without temporal constraint (Ours-None-ft) is inferior to the one with the soft constraint (Ours-Soft-ft) in all cases except the occluded regions (OCC) on MPI Sintel Final. On KITTI 2015, the improvements are marginal due to dominating complex motions. We conclude that fine-tuning with the soft constraint is in general beneficial even when complex motions violate the constant velocity assumption.

Results with Fine-tuning: Our soft constraint model fine-tuned on MPI Sintel (Ours-Soft-ft) achieves an EPE of 3.89 and 5.52 on Clean and Final, hence

outperforming all other unsupervised methods while even achieving comparable results to FlowNet fine-tuned on MPI Sintel Clean. Similarly on the test set[1], we outperform all other unsupervised methods with 7.23 and 8.81 EPE on Clean and Final, performing on par with supervised methods without fine-tuning, e.g. FlowNet and SPyNet. Fine-tuning on KITTI 2015 MV improves the performance to 6.59 in comparison to 8.10, the best achieved EPE by an unsupervised method so far. On the test set[1], we even achieve better performance than UnFlow that is trained on a large synthetic dataset (Synthia [10]) and KITTI Raw dataset. Figure 3 shows qualitative results of our fine-tuned models on each dataset. Compared to the two-frame formulation, our multi-frame formulation with occlusions results in more accurate optical flow fields with sharp motion discontinuities as well as occlusion estimates.

Table 3. Occlusion Estimates: We compare the performance of our occlusion estimation to other approaches on MPI Sintel and KITTI 2015 using the maximum F-Measure. Parentheses indicate cases where training was performed on the same dataset while ∗ marks cases where only the annotated samples were excluded from training. Note that S2D [48] is a supervised method.

Methods	MPI sintel clean	MPI sintel final	KITTI 2015
S2D [48]	-	**0.57**	-
MODOF [49]	-	**0.48**	-
OccAwareFlow-ft [19]	**(0.54)**	**(0.48)**	0.88*
Ours-Soft-ft	(0.49)	(0.44)	**0.91***

Occlusion Estimation: We evaluate our occlusion masks on both MPI Sintel and KITTI 2015 datasets. We compare our results quantitatively to OccAware-Flow [19], S2D [48], and MODOF [49] using the F-Measure (Table 3). While OccAwareFlow [19] obtains occlusion estimations considering the backward flow, S2D [48] uses a binary classification, and MODOF [49] uses a discrete-continuous optimization of an energy function.

With unsupervised fine-tuning on MPI Sintel (Ours-Soft-ft), we obtain comparable results to OccAwareFlow [19]. Learning occlusions on MPI Sintel in an unsupervised fashion is very difficult since occlusions often occur in untextured regions with limited guidance by the photometric loss. Even the supervised approach S2D struggles on the MPI Sintel dataset, only reaching a F-Measure of 0.57. Moreover, similar to the original PWC-Net [11], we estimate the optical flow and occlusion mask on quarter resolution. While larger occlusions are mostly estimated correctly, fine details are usually missing due to downsampling as can be observed in the qualitative results (Fig. 3). On KITTI 2015, the occlusion masks only contain pixels moving out of the image. Considering these masks,

---

[1] We submitted our results to MPI Sintel and KITTI 2015 under the name "Back2FutureFlow".

we reach the best performance with our unsupervised fine-tuned model (Outs-Soft-Kitti-ft). Note that several occlusions missing in the ground truth masks are correctly estimated by our method, e.g. the vehicles leaving the image in Fig. 3.

Contribution of RoamingImages: In contrast to other unsupervised approaches, we pre-train our model on our RoamingImages dataset since there are no simple multi-frame datasets available. This raises the question whether the reason for the success of our model is our dataset due to its size, simplicity or some other factor. To dispel this doubt, we pre-train UnFlow CSS [20] on our dataset and compare its performance to our pre-trained model. We use the code provided with default parameters only by changing the learning rate to $1e - 5$. As shown in Table 2, our pre-trained model (Ours-Hard) performs significantly better than UnFlow CSS trained on the same data (UnFlow-CSS (R)) on all datasets. This shows that the success of our approach is not solely based on our new dataset but critically depends on the proposed multi-frame formulation.

5   Conclusion

We presented a method for unsupervised learning of optical flow and occlusions from multiple frames. We proposed modifications to a state-of-the-art two-frame architecture for handling multiple frames in order to predict past and future optical flow as well as an occlusion map within a temporal window. We formulated unsupervised loss functions to exclude misleading information in occluded regions and incorporate a simple temporal model. In the experimental results, we show the benefits of the multi-frame formulation with occlusions over classical two-frame formulations and the importance of directly modeling temporal relations. We achieve impressive results with proper modeling and unsupervised training on a simple dataset outperforming all other unsupervised methods on complex datasets. After unsupervised fine-tuning, our approach is even on par with some fully supervised methods.

References

1. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artif. Intell. (AI) **17**(1–3), 185–203 (1981)
2. Black, M.J., Anandan, P.: A framework for the robust estimation of optical flow. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (1993)
3. Yamaguchi, K., McAllester, D., Urtasun, R.: Efficient Joint Segmentation, Occlusion Labeling, Stereo and Flow Estimation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 756–771. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_49
4. Yang, J., Li, H.: Dense, accurate optical flow estimation with piecewise parametric model. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)

5. Sun, D., Sudderth, E.B., Black, M.J.: Layered segmentation and optical flow estimation over time. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
6. Sevilla-Lara, L., Sun, D., Jampani, V., Black, M.J.: Optical flow with semantic segmentation and localized layers. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
7. Bai, M., Luo, W., Kundu, K., Urtasun, R.: Exploiting Semantic Information and Deep Matching for Optical Flow. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 154–170. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_10
8. Dosovitskiy, A., Fischer, P., Ilg, E., Haeusser, P., Hazirbas, C., Golkov, V., et al.: Flownet: learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
9. Ranjan, A., Black, M.: Optical flow estimation using a spatial pyramid network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
10. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: evolution of optical flow estimation with deep networks. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
11. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
12. Janai, J., Güney, F., Wulff, J., Black, M., Geiger, A.: Slow flow: exploiting high-speed cameras for accurate and diverse optical flow reference data. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
13. Mayer, N., Ilg, E., Haeusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., et al.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
14. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A Naturalistic Open Source Movie for Optical Flow Evaluation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 611–625. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33783-3_44
15. Yu, J.J., Harley, A.W., Derpanis, K.G.: Back to Basics: Unsupervised Learning of Optical Flow via Brightness Constancy and Motion Smoothness. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 3–10. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_1
16. Vijayanarasimhan, S., Ricco, S., Schmid, C., Sukthankar, R., Fragkiadaki, K.: Sfm-net: Learning of structure and motion from video. arXiv.org 1704.07804 (2017)
17. Pătrăucean, V., Handa, A., Cipolla, R.: Spatio-temporal video autoencoder with differentiable memory. In: Proceedings of the International Conference on Learning Representations (ICLR) (2016)
18. Ren, Z., Yan, J., Ni, B., Liu, B., Yang, X., Zha, H.: Unsupervised deep learning for optical flow estimation. In: Proceedings of the Conference on Artificial Intelligence (AAAI) (2017)
19. Wang, Y., Yang, Y., Yang, Z., Zhao, L., Wang, P., Xu, W.: Occlusion aware unsupervised learning of optical flow. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
20. Meister, S., Hur, J., Roth, S.: Unflow: unsupervised learning of optical flow with a bidirectional census loss. In: Proceedings of the Conference on Artificial Intelligence (AAAI) (2018)

21. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. Int. J. Robot. Res. (IJRR) **32**(11), 1231–1237 (2013)
22. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
23. Heeger, D.J.: Optical flow using spatiotemporal filters. Int. J. Comput. Vis. (IJCV) **1**(4), 279–302 (1988)
24. Fleet, D.J., Jepson, A.D.: Computation of component image velocity from local phase information. Int. J. Comput. Vis. (IJCV) **5**(1), 77–104 (1990)
25. Weickert, J., Schnörr, C.: Variational optic flow computation with a spatio-temporal smoothness constraint. J. Math. Imaging Vis. (JMIV) **14**(3), 245–255 (2001)
26. Stoll, M., Volz, S., Bruhn, A.: Joint trilateral filtering for multiframe optical flow. In: Proceedings of IEEE International Conference on Image Processing (ICIP) (2013)
27. Zimmer, H., Bruhn, A., Weickert, J.: Optic flow in harmony. Int. J. Comput. Vis. (IJCV) **93**(3), 368–388 (2011)
28. Ralli, J., Díaz, J., Ros, E.: Spatial and temporal constraints in variational correspondence methods. Mach. Vis. Appl. (MVA) **24**(2), 275–287 (2013)
29. Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., Bischof, H.: Anisotropic Huber-L1 optical flow. In: Proceedings of the British Machine Vision Conference (BMVC) (2009)
30. Volz, S., Bruhn, A., Valgaerts, L., Zimmer, H.: Modeling temporal coherence for optical flow. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2011)
31. Salgado, A., Sánchez, J.: Temporal constraints in large optical flow. In: Proceedings of the International Conference on Computer Aided Systems Theory (EURO-CAST) (2007)
32. Sun, D., Sudderth, E.B., Black, M.J.: Layered image motion with explicit occlusions, temporal consistency, and depth ordering. In: Advances in Neural Information Processing Systems (NIPS) (2010)
33. Wang, C.M., Fan, K.C., Wang, C.T.: Estimating optical flow by integrating multi-frame information. J. Inf. Sci. Eng. (JISE) (2008)
34. Black, M.J., Anandan, P.: Robust dynamic motion estimation over time. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (1991)
35. Kennedy, R., Taylor, C.J.: Optical Flow with Geometric Occlusion Estimation and Fusion of Multiple Frames. In: Tai, X.-C., Bae, E., Chan, T.F., Lysaker, M. (eds.) EMMCVPR 2015. LNCS, vol. 8932, pp. 364–377. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-14612-6_27
36. Garg, R., B.G., V.K., Carneiro, G., Reid, I.: Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 740–756. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_45
37. Xie, J., Girshick, R., Farhadi, A.: Deep3D: Fully Automatic 2D-to-3D Video Conversion with Deep Convolutional Neural Networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 842–857. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_51
38. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

39. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
40. Agrawal, P., Carreira, J., Malik, J.: Learning to see by moving. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
41. Long, G., Kneip, L., Alvarez, J.M., Li, H., Zhang, X., Yu, Q.: Learning Image Matching by Simply Watching Video. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 434–450. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_26
42. Alletto, S., Abati, D., Calderara, S., Cucchiara, R., Rigazio, L.: Transflow: Unsupervised motion flow by joint geometric and pixel-level estimation. arXiv.org (2017)
43. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Advances in Neural Information Processing Systems (NIPS) (2015)
44. Bruhn, A., Weickert, J., Schnörr, C.: Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods. Int. J. Comput. Vis. (IJCV) **61**(3), 211–231 (2005)
45. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
46. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015)
47. Mayer, N., Ilg, E., Fischer, P., Hazirbas, C., Cremers, D., Dosovitskiy, A., et al.: What makes good synthetic training data for learning disparity and optical flow estimation? Int. J. Comput. Vis. (IJCV) (2018)
48. Leordeanu, M., Zanfir, A., Sminchisescu, C.: Locally affine sparse-to-dense matching for motion and occlusion estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2013)
49. Xu, L., Jia, J., Matsushita, Y.: Motion detail preserving optical flow estimation. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **34**(9) (2012) 1744–1757