# Structure-from-Motion-Aware PatchMatch for Adaptive Optical Flow Estimation

Daniel Maurer[1][0000−0002−3835−2138], Nico Marniok[2], Bastian Goldluecke[2][0000−0003−3427−4029], and Andrés Bruhn[1][0000−0003−0423−7411]

[1] Institute for Visualization and Interactive Systems, University of Stuttgart, Germany
[2] Computer Vision and Image Analysis Group, University of Konstanz, Germany
{maurer,bruhn}@vis.uni-stuttgart.de
{nico.marniok,bastian.goldluecke}@uni-konstanz.de

**Abstract.** Many recent energy-based methods for optical flow estimation rely on a good initialization that is typically provided by some kind of feature matching. So far, however, these initial matching approaches are rather general: They do not incorporate any additional information that could help to improve the accuracy or the robustness of the estimation. In particular, they do not exploit potential cues on the camera poses and the thereby induced rigid motion of the scene. In the present paper, we tackle this problem. To this end, we propose a novel structure-from-motion-aware PatchMatch approach that, in contrast to existing matching techniques, combines two hierarchical feature matching methods: a recent two-frame PatchMatch approach for optical flow estimation (general motion) and a specifically tailored three-frame PatchMatch approach for rigid scene reconstruction (SfM). While the motion PatchMatch serves as baseline with good accuracy, the SfM counterpart takes over at occlusions and other regions with insufficient information. Experiments with our novel SfM-aware PatchMatch approach demonstrate its usefulness. They not only show excellent results for all major benchmarks (KITTI 2012/2015, MPI Sintel), but also improvements up to 50% compared to a PatchMatch approach without structure information.

## 1 Introduction

Since almost four decades the estimation of optical flow from image sequences is one of the most challenging tasks in computer vision. Despite of the recent success of learning-based approaches [2, 9, 18, 36, 23], global energy-based methods are still among the most accurate techniques for solving this task [16, 17, 22, 44]. Even if combined with partial learning [1, 33, 41, 42] such methods offer the advantage that they allow for a transparent modeling, since assumptions are explicitly stated in the underlying energy functional. However, since the complexity of the models has significantly grown within the last few years – recent methods try to estimate segmentation [33, 41, 44], occlusions [17, 44] or illumination changes [8] jointly with the optical flow – the minimization of the resulting non-convex energies has become an increasingly challenging problem.

In this context, many energy-based approaches [14, 22, 33, 41] rely on a suitable initialization provided by other methods. Among the most popular approaches that are considered useful as initialization are EpicFlow [30], Coarse-to-fine PatchMatch [15] and DiscreteFlow [25] – approaches that rely on the interpolation or fusion of *feature matches*. This has two main reasons: On the one hand, feature matching approaches are known to provide good results in the context of large displacements. On the other hand, they are typically based on some kind of filtering or a-posteriori regularization which renders the initialization sufficiently smooth and outlier-free. As a consequence, the initial flow field offers already a reasonable quality and the energy minimization starts with a good solution and is hence less likely to end up in undesired local minima.

While recent methods promote the use of feature-based approaches for initialization, they also show that integrating *additional information* in the estimation can be highly beneficial w.r.t. both accuracy and robustness [1, 16, 17, 33, 41]. Apart from considering domain-dependent semantic information [1, 5, 16, 33], it has proven useful to integrate structure constraints and symmetry cues. For instance, [41] proposed a method that jointly estimates the rigidity of each pixel together with its optical flow. Thereby structure constraints are imposed only on rigid parts of the scene. In contrast, [17] suggested an approach that exploits symmetry and consistency cues to jointly estimate forward and backward flows. This in turn, allows to infer occlusion information together with the optical flow.

Given the fact that the two aforementioned approaches as well as many other recent methods from the literature rely on a suitable initialization from feature-based methods, it is surprising that such information has *hardly entered* the initial feature matching step so far. While symmetry and consistency cues are at least considered in terms of simple forward-backward checks to detect occlusions and remove the corresponding outliers [9, 15, 30], structure constraints in terms of a rigid background motion have not found their way into feature matching approaches for computing the optical flow at all. Hence, it would be desirable to develop a feature-based method that allows to exploit structure information while still being able to estimate independently moving objects at the same time.

**Contributions.** In our paper, we develop such a hybrid method. In this context, our contributions are threefold. (i) First, we introduce a coarse-to-fine three-frame PatchMatch approach for estimating structure matches (SfM) that combines a depth-driven parametrization with different temporal selection strategies. While the parametrization robustifies the estimation by reducing the search space, the hierarchical optimization and the temporal selection improve the accuracy. (ii) Second, we propose a consistency-based selection scheme for combining matches from this structure-based PatchMatch approach and an unconstrained PatchMatch approach. Thereby, the backward flow allows us to identify reliable structure matches, while a robust voting scheme decides on the remaining cases. (iii) Finally, we embed the resulting matches into a full estimation pipeline. Using recent approaches for interpolation and refinement, our method provides dense results with sub-pixel accuracy. Experiments on all major benchmarks demonstrate the benefits of our novel SfM-aware PatchMatch approach.

### 1.1 Related Work

As mentioned, integrating additional information can render the estimation of the optical flow significantly more accurate and robust. We first comment on related work regarding the integration of such information, while afterwards we focus on related PatchMatch approaches for optical flow and scene structure.

**Rigid Motion.** In order to improve accuracy and robustness in case of a rigid background, one may enforce geometric assumptions such as the epipolar constraint [29, 38, 43, 44]. However, if this assumption is forced to hold for the entire scene, as proposed by Oisel *et al.* [29] and Yamaguchi *et al.* [43, 44], the approach is only applicable to fully rigid scenes, e.g. to those of the KITTI 2012 benchmark [11]. Although this problem can be slightly alleviated by soft constraints as proposed by Valgaerts *et al.* [37, 38], results for non-rigid scenes are typically not good. Hence, Wedel *et al.* [40] suggested to turn off the epipolar constraint for sequences with independent object motion. This, however, does not allow to exploit rigid body priors at all in the standard optical flow setting. Consequently, Gerlich and Eriksson [12] presented a more advanced approach that segments the scene into different regions with independent rigid body motions. While this strategy allows to handle automotive scenes with other rigdly moving objects quite well, e.g. sequences similar to the KITTI 2015 benchmark [24], it cannot model any type of non-rigid motion, e.g. as required for the different characters in the MPI Sintel benchmark [7]. In contrast, our SfM-aware PatchMatch approach combines information from general and SfM-based motion estimation. Hence, it is not restricted to fully rigid or object-wise rigid scenes.

**Mostly Rigid Motion.** Compared to [12], Wulff *et al.* [41] went a step further. Instead of requiring the scene to be object-wise rigid they assume the scene to be only mostly rigid. To this end, they suggested a complex iterative model that jointly segments the scene into foreground and background using semantic information as well as motion and structure cues while estimating the background motion with a dedicated epipolar stereo algorithm. In contrast to this approach, that uses the general optical flow method [25] as initialization and adaptively integrates strong rigidity priors later on in the estimation, our SfM-aware PatchMatch approach aims at integrating such priors already in the estimation of feature matches at the *very beginning* of the estimation – and this without the use of semantic information. Hence, our results are relevant for all methods relying on a suitable initialization – including the work of Wulff *et al.* [41] and other recent methods such as [17] or [33].

**Parametrized Models.** An alternative strategy that recently became very popular is to refrain from using global or object-wise rigidity priors and to model motions that are pixel- or piecewise rigid. Typically this is done by means of a suitable flow (over-)parametrization; see e.g. [13, 16, 24, 28, 39, 45]. For instance, Hornaček *et al.* [13] proposed a 9 DoF flow parametrization that models a locally rigid motion of planes. Similar, Yang *et al.* [45] and Hur and Roth [16, 17] suggested approaches that use a spatially coherent 8 DoF homography based on superpixels. In contrast to those methods, our SfM-aware PatchMatch approach

does not explicitly rely on an over-parametrization. Vice versa, it gains robustness by restricting the search space to 1D when calculating the SfM matches. Moreover, it estimates the flow pixel-wise instead of segment-wise. Hence, it is more suitable for general scenes with non-rigid motion and fine motion details.

**Semantic Information.** Another way to improve the accuracy and the robustness of the estimation is to consider semantic. For instance, Bai *et al.* [1] proposed to use instance-level segmentation to identify independently moving traffic participants before computing separate rigid motions for both the background and the participants. Similarly, Hur and Roth [16] make use of a CNN to integrate semantic information into a joint approach for estimating the flow and a temporally consistent semantic segmentation. Furthermore, Sevilla-Lara *et al.* [33] suggested a layered approach that relies on semantic information when switching between different motion models. Finally, there is also the method of Wulff *et al.* [41] (see mostly rigid motion). While semantic information often improves the results, it has to be particularly adapted to the given domain. As a consequence, the corresponding approaches do typically not generalize well across different applications or benchmarks. Hence, we do not rely on such information.

**PatchMatch.** In the context of unconstrained matching (optical flow), PatchMatch has been originally proposed by Barnes *et al.* [4]. Recent developments include the work of Bao *et al.* [3] that introduces an edge-preserving weighting scheme as well as the approach of Hu *et al.* [15] that improves accuracy and speed with a hierarchical matching strategy. Moreover, Gadot and Wolf [9] and Bailer *et al.* [2], have recently shown that feature learning can be beneficial. Despite of all the progress, however, none of the aforementioned optical flow methods includes structure information. In contrast, our SfM-aware approach exploits such information by explicitly using feature matches from a specifically tailored three-view stereo/SfM PatchMatch method.Also in the stereo/SfM context, there exists a vast literature on PatchMatch algorithms. There, PatchMatch has been first introduced by Bleyer et al. [6] who proposed a plane-fitting variant for the rectified case. Recent developments include the approaches of Shen [34] and Galliani *et al.* [10] who extended PatchMatch to the non-rectified two-view and multi-view case, respectively; see also [32, 46]. In contrast to all those methods, our SfM-aware PatchMatch approach not only extracts pure stereo information. Instead, it combines information from optical flow and stereo and is hence also applicable to non-rigid scenes with independent object motion. Moreover, it relies on a hierarchical optimization [15] which has not been used in the context of PatchMatch stereo so far. Finally, the SfM part of our algorithm uses a direct depth-parametrization. This, in turn, makes both the estimation very robust.

## 2 Method Overview

Let us start by giving a brief overview over the proposed method. As many recent optical flow techniques it relies on a multi-stage approach which includes steps for computing and refining an initial flow field; see e.g. [14, 17, 22, 33, 41]. However, in

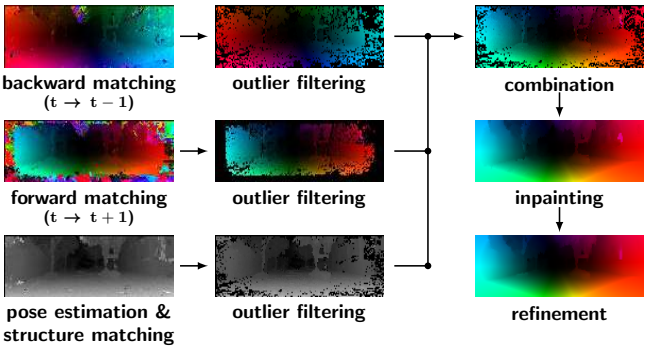


**Fig. 1.** Schematic overview over our SfM-aware PatchMatch approach.

contrast to most of these approaches that typically aim at improving an already given flow field, our method focuses on the generation of an accurate and robust initial flow field itself. To achieve this goal, our method integrates structure information into the feature matching process, which plays an essential role for the initialization [15, 25, 30]. This integration is motivated by the observation that many sequences contain a significant amount of rigid motion induced by the ego-motion of the camera [41]. Since this motion is constrained by the underlying stereo geometry, structure information can significantly improve the estimation.

In our multi-stage method, we realize this integration by combining two hierarchical feature matching approaches that complement each other: On the one hand, we use a recent two-frame PatchMatch approach for optical flow estimation [15]. This allows our method to estimate the unconstrained motion in the scene (forward and backward matches). On the other hand, we rely on a specifically tailored three-frame stereo/SfM PatchMatch approach (see Sec. 3) with preceding pose estimation [26]. This in turn, allows us our method to compute the rigid motion of the scene induced by the moving camera (structure matches). In order to discard outliers and combine the remaining matches, we perform a filtering approach for all matches followed by a consistency-based selection (see Sec. 4). Finally, we inpaint and refine the combined matches using recent methods from the literature [14, 22]. An overview of the entire approach is given in Fig. 1.

## 3 Structure Matching

In this section, we present our structure matching framework which builds upon the PatchMatch algorithm [4] – a randomized, iterative algorithm for approximate patch matching. In this context, we adopt ideas of the recently proposed Coarse-to-fine PatchMatch (CPM) for optical flow [15] and apply them in the context stereo/SfM estimation that relies on a depth-based parametrization [10, 31]. This not only enables the straightforward integration of multiple frames, but also allows to consider the concepts of temporal averaging and temporal selection [19], the latter one being a strategy for implicit occlusion handling.

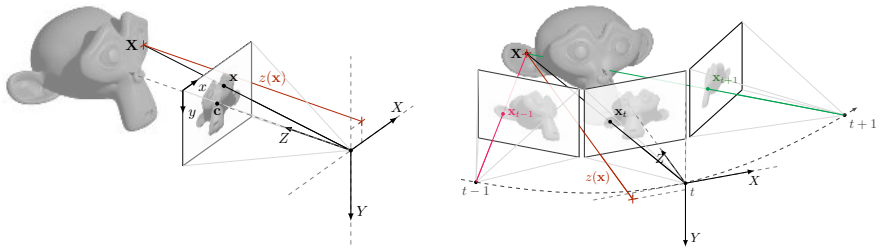


**Fig. 2. Left:** Illustration of the employed depth parametrization. **Right:** Illustration of corresponding points defined by the image location $\mathbf{x}_t$ and the associated depth value $z(\mathbf{x}_t)$. In this case, the 3D point is occluded in one view and could be handled with the idea of temporal selection. i.e. by the view from the other time step.

### 3.1 Depth-Based Parametrization

Let us start by deriving the employed depth-based parametrization. To this end, we assume that all images are captured by a calibrated perspective camera that possibly moves in space, i.e. the corresponding projection matrices $P_t = K\,[R_t|\mathbf{t}_t]$ are known. Here $R_t$ is a $3 \times 3$ rotation matrix and $\mathbf{t}_t$ a translation 3-vector that together describe the pose of the camera at a certain time step $t$. In addition, the $3 \times 3$ matrix $K$ denotes the intrinsic camera calibration matrix given by

$$K = \begin{pmatrix} s_x & 0 & c_x \\ 0 & s_y & c_y \\ 0 & 0 & 1 \end{pmatrix} , \tag{1}$$

where $(s_x, s_y)$ denotes the scaled focal length and $\mathbf{c} = (c_x, c_y)^\top$ denotes the principal point offset. Given the projection matrix $P_t$, a 3D point $\mathbf{X} \in \mathbb{R}^3$ is projected onto a 2D point $\mathbf{x} \in \mathbb{R}^2$ on the image plane by $\mathbf{x} = \pi(P_t\tilde{\mathbf{X}})$, where the tilde denotes homogeneous coordinates, such that

$$\tilde{\mathbf{X}} = \left(\mathbf{X}^\top, 1\right)^\top , \tag{2}$$

and $\pi$ maps a homogeneous coordinate $\tilde{\mathbf{x}}$ to its Euclidean counterpart $\mathbf{x}$

$$\pi(\tilde{\mathbf{x}}) = \begin{pmatrix} \tilde{x}_1/\tilde{x}_3 \\ \tilde{x}_2/\tilde{x}_3 \end{pmatrix} , \quad \text{with} \quad \tilde{\mathbf{x}} = \left(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3\right)^\top . \tag{3}$$

Now, to define our parametrization, we assume w.l.o.g. that the camera pose of the reference camera, i.e. the camera associated with the image taken at time $t$, is aligned with the world coordinate system and invert the previous described projection to specify a 3D point on the surface $\mathbf{s}$ by an image location $\mathbf{x}$ and the corresponding depth $z(\mathbf{x})$ along the optical axis; see Fig. 2. This leads to

$$\mathbf{X} = \mathbf{s}(\mathbf{x}, z(\mathbf{x})) = z(\mathbf{x})K^{-1}\tilde{\mathbf{x}}\,, \tag{4}$$

which allows us to describe correspondences throughout multiple images with a single unknown, the depth $z(\mathbf{x})$, by projecting onto the respective image planes

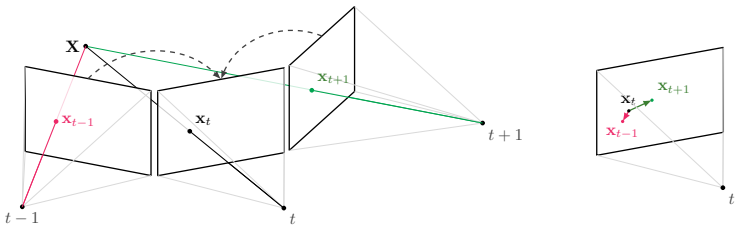


**Fig. 3.** Illustration showing the conversion procedure from a 3D point to the displacement vectors w.r.t. to the forward frame $t+1$ and backward frame $t-1$.

using the corresponding projection matrices; see Fig. 2. Finally, given three frames as in our case, with projection matrices $P_{t+1}$, $P_t$, and $P_{t-1}$, one can directly convert the estimated depth values to the corresponding displacement vectors w.r.t. to the forward frame $t+1$ and the backward frame $t-1$ (Fig. 3):

$$\mathbf{u}_{\text{st,fw}}(\mathbf{x}, z(\mathbf{x})) = \pi(P_{t+1}\tilde{\mathbf{s}}(\mathbf{x}, z(\mathbf{x})) - \pi(P_t\tilde{\mathbf{s}}(\mathbf{x}, z(\mathbf{x}))\,, \tag{5}$$

$$\mathbf{u}_{\text{st,bw}}(\mathbf{x}, z(\mathbf{x})) = \pi(P_{t-1}\tilde{\mathbf{s}}(\mathbf{x}, z(\mathbf{x})) - \pi(P_t\tilde{\mathbf{s}}(\mathbf{x}, z(\mathbf{x}))\,. \tag{6}$$

### 3.2 Hierarchical Matching

With the depth parametrization at hand we now turn to the actual matching. While applying the classical PatchMatch approach [4] directly to the problem typically yields noisy results due to non-existent explicit regularization, we resort to the idea of integrating a hierarchical coarse-to-fine scheme, which has shown to be less prone to noise in the context of optical flow estimation [15].

As in [15] we do not estimate the unknowns for all pixel locations, but for multiple collections of seeds $\mathcal{S}^l = \{s_m^l\}$ that are defined on each resolution level $l \in \{0, 1, \ldots, k-1\}$ of the coarse-to-fine pyramid. While the number of seeds remains the same for each resolution level, their spatial locations are given by

$$\mathbf{x}(s_m^l) = \lfloor \eta \cdot \mathbf{x}(s_m^{l-1}) \rceil \quad \text{for} \quad l \geq 1\,, \tag{7}$$

where $\lfloor \cdot \rceil$ is a function that returns the nearest integer value and $\eta = 0.5$ is the employed downsampling factor between two consecutive pyramid levels. Furthermore, the locations for $l = 0$ (full image resolution) are located at the cross points of a regular image grid with a spacing of 3 pixels and come with the default neighborhood system, defined via the spatial adjacency. In addition, these neighborhood relations remain fixed throughout the coarse-to-fine pyramid.

The matching is now performed in the classical coarse-to-fine manner: Starting at the coarsest resolution, each level is processed by iteratively performing a random search and a neighborhood propagation as in [4]. While the coarsest level uses a random initialization of the unknown depth, the subsequent levels are initialized with the depth values of the corresponding seeds of the next coarser level. Furthermore, the search radius for the random sampling is reduced exponentially throughout the coarse-to-fine pyramid, such that the random search is restricted to values near the current best depth estimate.

### 3.3 Cost Computation and Temporal Averaging / Selection

Since we consider three images, there are several possibilities how to compute the matching cost between corresponding patches. One possible choice is to compute all pairwise similarity measures w.r.t. the reference patch and average the costs. While this renders the estimation more robust if the actual 3D point is visible in all views, it may lead to deteriorated results in case of occlusions. In order to deal with this, one can apply the idea of temporal selection [19] and compute all pairwise similarity measures w.r.t. the reference patch, but only consider the lowest pairwise cost as overall cost. Thereby it can be ensured that, as long as the reference patch can be found in at least one view and is occluded in the remaining ones, the correct correspondence retains a small cost. In our experiments we will use both approaches, temporal averaging and temporal selection.

Finally, we utilize SIFT descriptors [15, 20, 21] in order to compute the similarity between two corresponding locations. This also renders the matching more robust than operating directly on the intensity values. Regarding the cost function we follow [15] and apply a robust $L^1$-loss. The resulting forward and backward structure matching costs $C_{t+1}$ and $C_{t-1}$ are then given by

$$C_{t+1}(\mathbf{x}, z(\mathbf{x})) = ||\mathbf{f}_{\mathrm{SIFT}}(\pi(P_{t+1}\tilde{\mathbf{s}}(\mathbf{x}, z(\mathbf{x}))) - \mathbf{f}_{\mathrm{SIFT}}(\pi(P_t\tilde{\mathbf{s}}(\mathbf{x}, z(\mathbf{x}))||_1 \ , \tag{8}$$

$$C_{t-1}(\mathbf{x}, z(\mathbf{x})) = ||\mathbf{f}_{\mathrm{SIFT}}(\pi(P_{t-1}\tilde{\mathbf{s}}(\mathbf{x}, z(\mathbf{x}))) - \mathbf{f}_{\mathrm{SIFT}}(\pi(P_t\tilde{\mathbf{s}}(\mathbf{x}, z(\mathbf{x}))||_1 \ , \tag{9}$$

where $\mathbf{f}_{\mathrm{SIFT}}$ denotes the SIFT-feature and $||\cdot||_1$ is the $L^1$-norm. The corresponding temporal averaging and temporal selection costs read

$$C_{\mathrm{avg}}(\mathbf{x}, z(\mathbf{x})) = \tfrac{1}{2}(C_{t+1}(\mathbf{x}, z(\mathbf{x})) + C_{t-1}(\mathbf{x}, z(\mathbf{x}))) \ , \tag{10}$$

$$C_{\mathrm{ts}}(\mathbf{x}, z(\mathbf{x})) = \min(C_{t+1}(\mathbf{x}, z(\mathbf{x})), C_{t-1}(\mathbf{x}, z(\mathbf{x}))) \ . \tag{11}$$

### 3.4 Outlier Handling

Finally, we extend the classical bi-directional consistency check to our three-view setting. Therefore, we not only estimate the depth values with frame $t$ as reference view but also with the other two frames as reference. Then we take the estimated depth value $z_t(\mathbf{x})$ at frame $t$, project it into the frames $t+1$ and $t-1$, take the estimated depth values $z_{t+1}(\mathbf{x})$ and $z_{t-1}(\mathbf{x})$ there, and project them back to frame $t$. Only if at least one of the two backprojections maps to the starting point $\mathbf{x}$, the depth value $z_t(\mathbf{x})$ is considered valid. In this case, the forward/backward structure matches can be computed from $z_t(\mathbf{x})$ via Eqs. (5)-(6).

## 4 Combining Matches

At this point, we have computed filtered forward and backward structure matches from frame $t$ to frames $t+1$ and $t-1$. For the sake of clarity let us denote these matches by $\hat{\mathbf{u}}_{\mathrm{st,fw}}$ and $\hat{\mathbf{u}}_{\mathrm{st,bw}}$. Moreover, as indicated in Fig. 1. we have also computed the corresponding forward and backward optical flow matches between the

same frames with a hierarchical PatchMatch approach for unconstrained motion [15]. Since these optical flow matches underwent a classical bi-directional consistency check to remove outliers (which requires to additionally compute matches from frames $t+1$ and $t-1$ to frame $t$), let us denote them by $\hat{\mathbf{u}}_{\text{of,fw}}$ and $\hat{\mathbf{u}}_{\text{of,bw}}$.

The goal of the combination step is now to fuse these four matches in such a way such that rigid parts of the scene can benefit from the structure matches. Thereby one has to keep in mind that optical flow matches may explain rigid motion, while structure matches are typically wrong in the context of independent object motion. To avoid using structural matches at inappropriate locations, we propose a conservative approach: We augment the optical flow matches with the matches obtained from the structure matching. This means that we always keep the match of the forward flow, if it has passed the outlier filtering. Otherwise, however, we consider to augment the final matches at this location by the match of the structure matching approach. In order to decide if such a structure match should really be considered, we propose three different approaches (see Fig. 4):

**Permissive Approach.** The first approach is the most permissive approach. It includes all structure matches $\hat{\mathbf{u}}_{\text{st,fw}}$ that have passed the outlier filtering at locations where no forward optical flow match $\hat{\mathbf{u}}_{\text{of,fw}}$ is available.

**Restrictive Approach.** The second approach is more restrictive. Instead of including all structure matches, we enforce an additional consistency check. This allows to reduce the probability of blindly including possibly false matches. For this consistency check we make use of the backward optical flow match $\hat{\mathbf{u}}_{\text{of,bw}}$. We only consider the forward structure match $\hat{\mathbf{u}}_{\text{st,fw}}$, if its backward variant $\hat{\mathbf{u}}_{\text{st,bw}}$ is consistent with the backward optical flow match $\hat{\mathbf{u}}_{\text{of,bw}}$. In case the additional consistency check cannot be performed, because the backward optical flow match did not pass the outlier filtering, we do not consider the structure match.

**Voting Approach.** Finally, we propose a voting approach that enforces the additional consistency check as in the restrictive approach but still allows to include structure matches in case the additional consistency check cannot be performed. The decision if such non-checkable structure matches should be included is conducted for each sequence separately. It is based on a voting scheme: All locations, that contain a valid match for the forward, backward and structure match are eligible to vote. If the structure match is consistent with both the forward and the backward match, we count this as a vote in favor of including non-checkable matches. If the votes surpass a certain threshold (80% in our experiments) all non-checkable structure matches are added. This can be seen as a detection scheme that allows to identify scenes with a large amount of ego-motion.

## 5 Evaluation

**Evaluation Setup.** In order to evaluate our new approach, we used the following components within our pipeline (cf. Fig. 1): The pose estimation uses the OpenMVG [27] implementation of the incremental SfM approach [26], the forward and backward matching employ the Coarse-to-fine PatchMatch (CPM) [15]

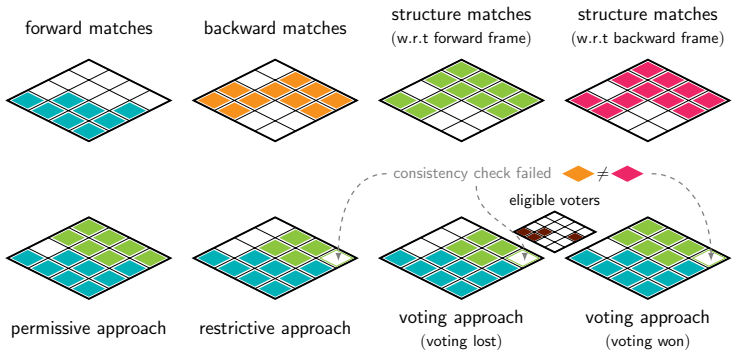


**Fig. 4.** Illustration showing the different strategies to combine the computed matches. **Top:** Color coded input matches. White denotes no match. **Bottom:** Fusion results.

approach, the structure matching and consistent combination are performed as described in Sec. 3 and 4, respectively, followed by a robust interpolation of the combined correspondences (RIC) using [14]. Finally, the inpainted matches are refined using the order-adaptive illumination-aware refinement method (OIR) [22]. Except for the refinement, where we optimized [35] the three weighting parameters per benchmark using the training data, we used the default parameters.

**Benchmarks.** To evaluate the performance of our approach, we consider three different benchmarks: the KITTI 2012 [11], the KITTI 2015 [24], and the MPI Sintel [7] benchmark. These benchmarks exhibit an increasing amount of ego-motion induced optical flow. While KITTI 2012 consists of pure ego-motion, KITTI 2015 additionally includes motion of other traffic participants. Finally, MPI Sintel also contains non-rigid motion from animated characters.

**Baseline.** To measure improvements, we establish a baseline that does not use structure information and only relies on forward optical flow matches (CPM). As Tab. 1 shows, our baseline outperforms most of the related approaches. Only DF+OIR [22] performs slightly better, due to the advanced DF matches [25].

**Structure Matching.** Next, we investigate the performance of our novel structure matching approach on its own. Therefore, we replace the matching approach (CPM) in our baseline with three variants of our structure matching approach (CPMz): a two-frame variant, a three-frame variant with temporal averaging and a three-frame variant with temporal selection. As the results in Tab. 1 show, structure matching significantly outperforms the baseline in pure ego-motion scenes, while it naturally has problems in scenes with independent motion. Moreover, they show that the use of multiple frames pays off. However, while for the KITTI benchmarks the robustness of temporal averaging is more beneficial than the occlusion handling of temporal selection, the opposite holds for the MPI Sintel benchmark. This, in turn, might be attributed to the fact that MPI Sintel contains a larger amount of occlusions. Since both strategies have their advantages, we consider both variants for our further evaluation.

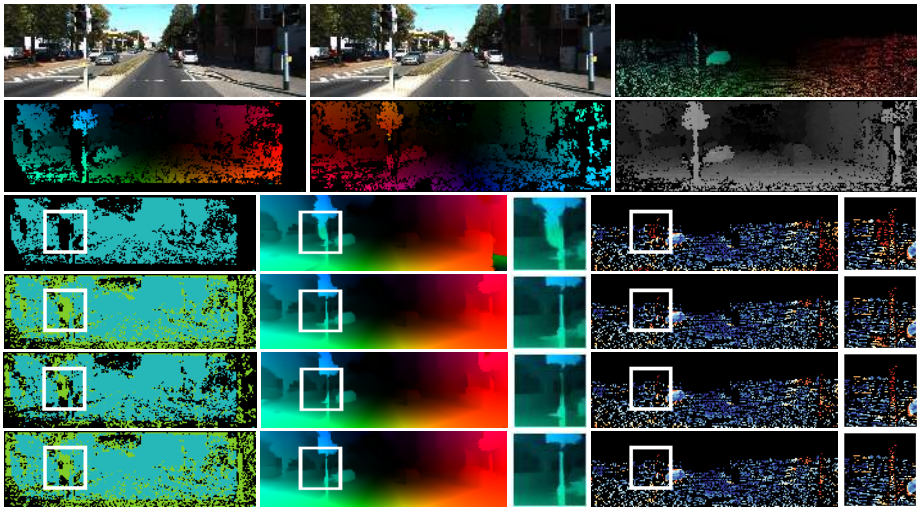

**Fig. 5.** Example for the KITTI 2015 benchmark [24] (#186). **First row:** Reference frame, subsequent frame, ground truth. **Second row:** Forward matches, structure matches (depth visualization). **Following rows. From left to right:** Used matches (color-coding see Fig. 4), final result, bad pixel visualization. **From top to bottom:** Baseline, permissive approach, restrictive approach, voting approach.

**Unconstrained Matching.** Apart from the baseline we also evaluated two additional variants solely based on unconstrained matching: a variant only using backward matches and a variant that augments the forward matches with backward matches. To this end, we assume a constant motion model, i.e. $\hat{\mathbf{u}}_{\text{of,fw}} = -\hat{\mathbf{u}}_{\text{of,bw}}$. The results for the backward flow in Tab. 1 show that such a simple model does not allow to leverage useful information to predict the forward flow. Even the augmented variant does not improve compared to the baseline.

**Combined Approach.** Let us now turn towards the evaluation of our combined approach. In this context, we compare the impact of the different combination strategies. As one can see in Tab. 1, the permissive approach is not an option. While it works well for dominating ego-motion, it includes too many false structure matches in case of independent object motion. In contrast, the restrictive approach prevents the inclusion of false structure matches, but cannot make use of the full potential of such matches in scenes with dominating ego-motion. Nevertheless, it already outperforms the baseline significantly and gives the best results for MPI Sintel. Finally, the voting approach combines the advantages of both schemes. It yields the best results for KITTI 2012/2015 with improvements up to 50% compared to the baseline, while still offering an improvement w.r.t. MPI Sintel. This observation is also confirmed by the examples in Fig. 5/6. They show the usefulness of including structure matches in occluded areas and the importance of filtering false structure matches in general.

**Table 1.** Results for the training datasets of the KITTI 2012 [11] (all pixels), KITTI 2015 [24] (all pixels) and the MPI Sintel [7] benchmarks (clean render path) in terms of the average endpoint error (AEE) and the percentage of bad pixels (BP, 3px threshold).

| **method** | | | | **KITTI 2012** | | **KITTI 2015** | | **Sintel** |
|---|---|---|---|---|---|---|---|---|
| name | matching | inpainting | refinement | AEE | BP | AEE | BP | AEE |
| **related approaches (+ baseline)** | | | | | | | | |
| CPM-Flow [15] | CPM | EPIC | EPIC | 3.00 | 14.58 | 7.78 | 22.86 | 2.00 |
| RIC-Flow [14] | CPM | RIC | OpenCV | 2.94 | 10.94 | 7.24 | 21.46 | 2.16 |
| CPM+OIR [22] | CPM | EPIC | OIR | 2.78 | 9.68 | 7.36 | 19.21 | 1.99 |
| DF+OIR [22] | DF | EPIC | OIR | **2.34** | 9.29 | **5.89** | **18.10** | **1.91** |
| baseline | CPM | RIC | OIR | 2.61 | **8.98** | 6.82 | 18.70 | 1.95 |
| **only structure matching** | | | | | | | | |
| two-frame | CPMz | RIC | OIR | 2.25 | 9.47 | 9.15 | 23.02 | 17.09 |
| temporal averaging | CPMz | RIC | OIR | **1.25** | **6.51** | **7.85** | **19.11** | 20.68 |
| temporal selection | CPMz | RIC | OIR | 1.43 | 6.69 | 8.06 | 19.52 | **15.69** |
| **only unconstrained matching** | | | | | | | | |
| backward flow | CPM | RIC | OIR | 6.90 | 43.96 | 11.57 | 44.12 | 4.00 |
| forward flow | CPM | RIC | OIR | **2.61** | **8.98** | **6.82** | **18.70** | **1.95** |
| combined fw&bw | CPM | RIC | OIR | 4.53 | 18.93 | 9.54 | 27.42 | 2.05 |
| **combined (temporal selection)** | | | | | | | | |
| permissive approach | CPM/CPMz | RIC | OIR | **1.47** | 5.91 | 4.95 | 14.12 | 2.53 |
| restrictive approach | CPM/CPMz | RIC | OIR | 1.60 | 6.22 | 5.20 | 15.10 | **1.88** |
| voting approach | CPM/CPMz | RIC | OIR | 1.48 | **5.82** | **4.91** | **13.95** | 1.90 |
| **combined (temporal averaging)** | | | | | | | | |
| permissive approach | CPM/CPMz | RIC | OIR | **1.30** | 5.71 | 4.21 | 13.72 | 2.92 |
| restrictive approach | CPM/CPMz | RIC | OIR | 1.59 | 6.17 | 5.04 | 14.97 | **1.90** |
| voting approach | CPM/CPMz | RIC | OIR | **1.30** | **5.67** | **4.16** | **13.61** | 1.92 |
| **recent literature** | | | | | | | | |
| PWC-Net [36] | CVPR '18 | | | 4.14 | – | 10.35 | 33.67 | 2.55 |
| FlowNet2 [18] | CVPR '18 | | | 4.09 | – | 10.06 | 30.37 | 2.02 |
| UnFlow [23] | AAAI '18 | | | **3.29** | – | **8.10** | 23.27 | – |
| DCFlow [42] | CVPR '17 | | | – | – | – | 15.09 | – |
| MR-Flow [41] | CVPR '17 | | | – | – | – | 14.09 | **1.83** |
| Mirror Flow [17] | ICCV '17 | | | – | – | – | **9.98** | – |
| **learning approaches (fine tuned)** | | | | | | | | |
| PWC-Net-ft[36] | CVPR '18 | | | (1.45) | – | (2.16) | (9.80) | (1.70) |
| FlowNet2-ft [18] | CVPR '17 | | | (1.28) | – | (2.30) | (8.61) | (1.45) |
| UnFlow-ft [23] | AAAI '18 | | | (1.14) | – | (1.86) | (7.40) | – |

**Comparison to the Literature.** Finally, we compare our method to other approaches from the literature. To this end, we consider both the training and the test data sets; see Tab. 1 and Tab. 2, respectively. Regarding the training data sets, our method generally yields better results than recent learning approaches without fine-tuning (PWC-Net [36], FlowNet2 [18], UnFlow [23]). Moreover, it also outperforms DCFlow [42] and MR-Flow [41] on the KITTI 2015 benchmark. Only MirrorFlow [17] (KITTI 2015) and MR-Flow (MPI Sintel) provide better results. This good performance holds for the test data sets as well, for which we

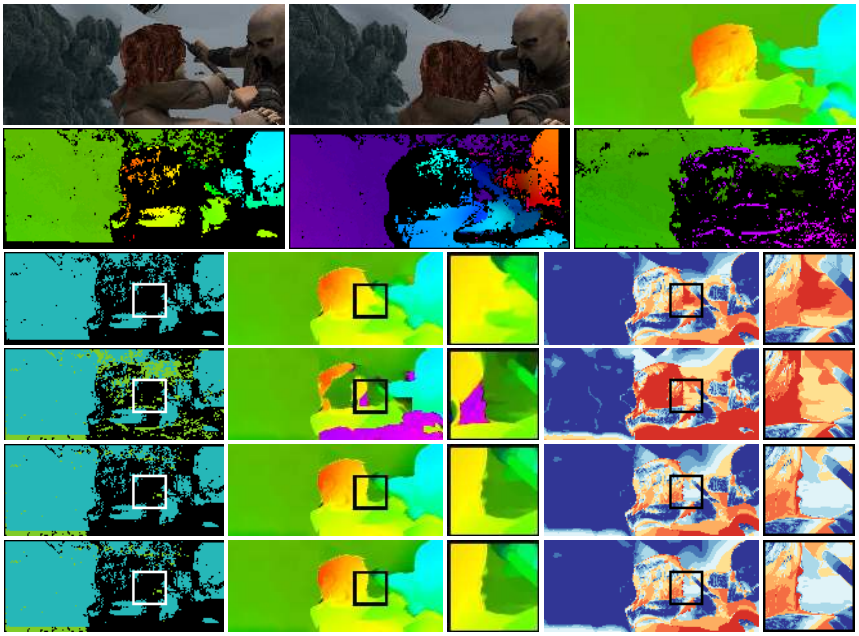

**Fig. 6.** Example for the MPI Sintel benchmark [7] (ambush5 #44). **First row:** Reference frame, subsequent frame, ground truth. **Second row:** Forward matches, structure matches (forward match visualization). **Following rows. From left to right:** Used matches (color-coding see Fig. 4), final result, bad pixel visualization. **From top to bottom:** Baseline, permissive approach, restrictive approach, voting approach.

evaluated the approaches that had performed best on the training data. Here, on KITTI 2012, our method performs favorably (all pixels) even compared to methods based on pure ego-motion and semantic information. Moreover, it also outperforms recent approaches with an explicit SfM background estimation (MR-Flow) on KITTI 2015. Finally, ranking second and sixth our method also yields an excellent performance on the clean and final set of MPI Sintel, respectively. This shows that our method not only works well in the context of pure ego-motion but can also handle a significant amount of independent object motion.

**Fixed Parameter Set.** Finally, we investigate how the results change when not optimizing the refinement parameters individually for each benchmark. To this end, we considered the voting approach with temporal averaging and conducted an experiment on the training data with *all parameters fixed*. As Tab. 3 shows the results hardly deteriorate when using a single parameter set for all benchmarks.

**Runtime.** The runtime of the pipeline excluding the pose estimation is 32s for one frame of size 1024×436 (MPI Sintel) using three cores on an Intel® Core™ i7-7820X CPU @ 3.6GHz, which splits into: 5.5s matching (incl. outlier filtering), <0.1s combination, 1.5s inpainting and 25s refinement. The pose estimation is run on the entire image sequence, which takes 83s for a sequence with 50 frames.

**Table 2.** Top 10 non-anonymous optical flow methods on the test data of the KITTI 2012/2015 [11, 24] and of the MPI Sintel benchmark [7], excluding scene flow methods.

| KITTI 2012 | Out-Noc | Out-All | Avg-Noc | Avg-All |
|---|---|---|---|---|
| SPS-Fl[1] | 3.38 % | 10.06 % | 0.9 px | 2.9 px |
| PCBP-Flow[1] | 3.64 % | 8.28 % | 0.9 px | 2.2 px |
| SDF[2] | 3.80 % | 7.69 % | 1.0 px | 2.3 px |
| MotionSLIC[1] | 3.91 % | 10.56 % | 0.9 px | 2.7 px |
| **our approach** | **4.02 %** | **6.15 %** | **1.0 px** | **1.5 px** |
| PWC-Net | 4.22 % | 8.10 % | 0.9 px | 1.7 px |
| UnFlow | 4.28 % | 8.42 % | 0.9 px | 1.7 px |
| MirrorFlow | 4.38 % | 8.20 % | 1.2 px | 2.6 px |
| ImpPB+SPCI | 4.65 % | 13.47 % | 1.1 px | 2.9 px |
| CNNF+PMBP | 4.70 % | 14.87 % | 1.1 px | 3.3 px |

| KITTI 2015 | Fl-bg | Fl-fg | Fl-all |
|---|---|---|---|
| PWC-Net | 9.66 % | 9.31 % | 9.60 % |
| MirrorFlow | 8.93 % | 17.07 % | 10.29 % |
| SDF[2] | 8.61 % | 23.01 % | 11.01 % |
| UnFlow | 10.15 % | 15.93 % | 11.11 % |
| CNNF+PMBP | 10.08 % | 18.56 % | 11.49 % |
| **our approach** | **9.66 %** | **22.73 %** | **11.83 %** |
| MR-Flow[2] | 10.13 % | 22.51 % | 12.19 % |
| DCFlow | 13.10 % | 23.70 % | 14.86 % |
| SOF[2] | 14.63 % | 22.83 % | 15.99 % |
| JFS[2] | 15.90 % | 19.31 % | 16.47 % |

| MPI Sintel clean | all | matched | unmatched |
|---|---|---|---|
| MR-Flow[2] | 2.527 | 0.954 | 15.365 |
| **our approach** | **2.910** | **1.016** | **18.357** |
| FlowFields+ | 3.102 | 0.820 | 21.718 |
| CPM2 | 3.253 | 0.980 | 21.812 |
| MirrorFlow | 3.316 | 1.338 | 19.470 |
| DF+OIR | 3.331 | 0.942 | 22.817 |
| S2F-IF | 3.500 | 0.988 | 23.986 |
| SPM-BPv2 | 3.515 | 1.020 | 23.865 |
| DCFlow | 3.537 | 1.103 | 23.394 |
| RicFlow | 3.550 | 1.264 | 22.220 |

| MPI Sintel final | all | matched | unmatched |
|---|---|---|---|
| PWC-Net | 5.042 | 2.445 | 26.221 |
| DCFlow | 5.119 | 2.283 | 28.228 |
| FlowFieldsCNN | 5.363 | 2.303 | 30.313 |
| MR-Flow[2] | 5.376 | 2.818 | 26.235 |
| S2F-IF | 5.417 | 2.549 | 28.795 |
| **our approach** | **5.466** | **2.683** | **28.147** |
| InterpoNet_ff | 5.535 | 2.372 | 31.296 |
| RicFlow | 5.620 | 2.765 | 28.907 |
| InterpoNet_cpm | 5.627 | 2.594 | 30.344 |
| ProbFlowFields | 5.696 | 2.545 | 31.371 |

[1] uses epipolar geometry as a hard constraint, only applicable to pure ego-motion
[2] exploits semantic information

**Table 3.** Impact of refinement parameter optimization.

| method | | KITTI 2012 | | KITTI 2015 | | Sintel |
|---|---|---|---|---|---|---|
| name | parameters | AEE | BP | AEE | BP | AEE |
| voting approach | individually optimized | **1.30** | **5.67** | **4.16** | **13.61** | **1.92** |
| voting approach | single parameter set | 1.31 | 5.70 | **4.16** | 13.70 | 1.93 |

## 6 Conclusion

In this paper, we addressed the problem of integrating structure information into feature matching approaches for computing the optical flow. To this end, we developed a hierarchical depth-parametrized three-frame SfM/stereo PatchMatch approach with temporal selection and preceding pose estimation. By adaptively combining the resulting matches with those of a recent PatchMatch approach for general motion estimation, we obtained a novel SfM-aware method that benefits from a global rigidity prior, while still being able to estimate independently moving objects. Experiments not only showed excellent results on all major benchmarks (KITTI 2012/2015, MPI Sintel), they also demonstrated consistent improvements over a baseline without structure information. Since our approach is based on inpainting and refining advanced feature matches, it offers another advantage: Other optical flow methods can easily benefit from it by incorporating its matches or the resulting dense flow fields as initialisation.

**Acknowledgments.** We thank the German Research Foundation (DFG) for financial support within projects B04 and B05 of SFB/Transregio 161.

## References


1. Bai, M., Luo, W., Kundu, K., Urtasun, R.: Exploiting semantic information and deep matching for optical flow. In: Proc. European Conference on Computer Vision. pp. 154–170 (2016)
2. Bailer, C., Varanasi, K., Stricker, D.: CNN-based patch matching for optical flow with thresholded Hinge embedding loss. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 2710–2719 (2017)
3. Bao, L., Yang, Q., Jin, H.: Fast edge-preserving PatchMatch for large displacement optical flow. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 1510–1517 (2014)
4. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch: A randomized correspondence algorithm for structural image editing. ACM Transactions on Graphics **28**(3), 24 (2009)
5. Behl, A., Jafari, O., Mustikovela, S., Alhaija, H., Rother, C., Geiger, A.: Bounding boxes, segmentations and object coordinates: how important is recognition for 3D scene flow estimation in autonomous driving scenarios? In: Proc. IEEE International Conference on Computer Vision. pp. 2574–2583 (2017)
6. Bleyer, M., Rhemann, C., Rother:, C.: PatchMatch stereo - stereo matching with slanted support windows. In: Proc. British Machine Vision Conference. pp. 14:1–14:11 (2011)
7. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Proc. European Conference on Computer Vision. pp. 611–625 (2012)
8. Demetz, O., Stoll, M., Volz, S., Weickert, J., Bruhn, A.: Learning brightness transfer functions for the joint recovery of illumination changes and optical flow. In: Proc. European Conference on Computer Vision. pp. 455–471 (2014)
9. Gadot, D., Wolf, L.: PatchBatch: a batch augmented loss for optical flow. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 4236–4245 (2016)
10. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: Proc. IEEE International Conference on Computer Vision. pp. 873–881 (2015)
11. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361 (2012)
12. Gerlich, T., Eriksson, J.: Optical flow for rigid multi-motion scenes. In: Proc. IEEE International Conference on 3D Vision. pp. 212–220 (2016)
13. Hornacek, M., Besse, F., Kautz, J., Fitzgibbon, A.W., Rother, C.: Highly overparameterized optical flow using PatchMatch belief propagation. In: Proc. European Conference on Computer Vision. pp. 220–234 (2014)
14. Hu, Y., Li, Y., Song, R.: Robust interpolation of correspondences for large displacement optical flow. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 481–489 (2017)
15. Hu, Y., Song, R., Li, Y.: Efficient Coarse-to-fine PatchMatch for large displacement optical flow. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 5704–5712 (2016)
16. Hur, J., Roth, S.: Joint optical flow and temporally consistent semantic segmentation. In: Proc. ECCV Workshop on Computer Vision for Road Scene Understanding and Autonomous Driving. pp. 163–177 (2016)

17. Hur, J., Roth, S.: MirrorFlow: exploiting symmetries in joint optical flow and occlusion estimation. In: Proc. IEEE International Conference on Computer Vision. pp. 312–321 (2017)
18. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: FlowNet 2.0: evolution of optical flow estimation with deep networks. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 1647–1655 (2017)
19. Kang, S.B., Szeliski, R., Chai, J.: Handling occlusions in dense multi-view stereo. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 103–110 (2001)
20. Liu, C., Yuen, J., Torralba, A.: SIFT flow: Dense correspondence across scenes and its applications. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(5), 978–994 (2011)
21. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
22. Maurer, D., Stoll, M., Bruhn, A.: Order-adaptive and illumination-aware variational optical flow refinement. In: Proc. British Machine Vision Conference. pp. 662:1–662:13 (2017)
23. Meister, S., Hur, J., Roth, S.: UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In: Proc. AAAI Conference on Artificial Intelligence (2018)
24. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 3061–3070 (2015)
25. Menze, M., Heipke, C., Geiger, A.: Discrete optimization for optical flow. In: Proc. German Conference on Pattern Recognition. pp. 16–28 (2015)
26. Moulon, P., Monasse, P., Marlet, R.: Adaptive structure from motion with a contrario model estimation. In: Proc. Asian Conference on Computer Vision. pp. 257–270 (2012)
27. Moulon, P., Monasse, P., Marlet, R., Others: OpenMVG. An Open Multiple View Geometry library. https://github.com/openMVG/openMVG
28. Nir, T., Bruckstein, A.M., Kimmel, R.: Over-parameterized variational optical flow. International Journal of Computer Vision **76**(2), 205–216 (2007)
29. Oisel, L., Memin, E., Morin, L., Labit, C.: Epipolar constrained motion estimation for reconstruction from video sequences. In: Proc. SPIE. vol. 3309, pp. 460–468 (1998)
30. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: Epicflow: Edge-preserving interpolation of correspondences for optical flow. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 1164–1172 (2015)
31. Robert, L., Deriche, R.: Dense depth map reconstruction: a minimization and regularization approach which preserves discontinuities. In: Proc. European Conference on Computer Vision. pp. 439–451 (1996)
32. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: Proc. European Conference on Computer Vision. pp. 501–518 (2016)
33. Sevilla-Lara, L., Sun, D., Jampani, V., Black, M.J.: Optical flow with semantic segmentation and localized layers. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 3889–3898 (2016)
34. Shen, S.: Accurate multiple view 3D reconstruction using patch-based stereo for large-scale scenes. IEEE Transactions on Image Processing **22**(5), 1901–1914 (2013)

35. Stoll, M., Volz, S., Maurer, D., Bruhn, A.: A time-efficient optimisation framework for parameters of optical flow methods. In: Proc. Scandinavian Conference on Image Analysis. pp. 41–53 (2017)
36. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (2018)
37. Valgaerts, L., Bruhn, A., Mainberger, M., Weickert, J.: Dense versus sparse approaches for estimating the fundamental matrix. International Journal of Computer Vision **96**(2), 212–234 (2012)
38. Valgaerts, L., Bruhn, A., Weickert, J.: A variational model for the joint recovery of the fundamental matrix and the optical flow. In: Proc. German Conference on Pattern Recognition. pp. 314–324 (2008)
39. Vogel, C., Schindler, K., Roth, S.: 3D scene flow estimation with a piecewise rigid scene model. International Journal of Computer Vision **115**(1), 1–28 (2015)
40. Wedel, A., Cremers, C., Pock, T., Bischof, H.: Structure-and motion-adaptive regularization for high accuracy optic flow. In: Proc. IEEE International Conference on Computer Vision. pp. 1663–1668 (2009)
41. Wulff, J., Sevilla-Lara, L., Black, M.J.: Optical flow in mostly rigid scenes. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 6911–6920 (2017)
42. Xu, J., Ranftl, R., Koltun, V.: Accurate optical flow via direct cost volume processing. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 5807–5815 (2017)
43. Yamaguchi, K., McAllester, D., Urtasun, R.: Robust monocular epipolar flow estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 1862–1869 (2013)
44. Yamaguchi, K., McAllester, D., Urtasun, R.: Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In: Proc. European Conference on Computer Vision. pp. 756–771 (2014)
45. Yang, J., Li, H.: Dense, accurate optical flow estimation with piecewise parametric model. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 1019–1027 (2015)
46. Zheng, E., Dunn, E., Jojic, V., Frahm, J.M.: PatchMatch based joint view selection and depthmap estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 1510–1517 (2014)