Object Scene Flow for Autonomous Vehicles

Moritz Menze
Leibniz Universität Hannover
menze@ipi.uni-hannover.de

Andreas Geiger
MPI Tübingen
andreas.geiger@tue.mpg.de

Abstract

This paper proposes a novel model and dataset for 3D scene flow estimation with an application to autonomous driving. Taking advantage of the fact that outdoor scenes often decompose into a small number of independently moving objects, we represent each element in the scene by its rigid motion parameters and each superpixel by a 3D plane as well as an index to the corresponding object. This minimal representation increases robustness and leads to a discrete-continuous CRF where the data term decomposes into pairwise potentials between superpixels and objects. Moreover, our model intrinsically segments the scene into its constituting dynamic components. We demonstrate the performance of our model on existing benchmarks as well as a novel realistic dataset with scene flow ground truth. We obtain this dataset by annotating 400 dynamic scenes from the KITTI raw data collection using detailed 3D CAD models for all vehicles in motion. Our experiments also reveal novel challenges which cannot be handled by existing methods.

1. Introduction

"Most of the structures in the visual world are rigid or at least nearly so." *David Marr [20]*

The estimation of dense 3D motion fields, widely termed *scene flow*, is currently gaining increasing attention. Dense motion vectors yield insights into the geometric layout of a scene as well as its decomposition into individually moving objects. Important applications include mobile robotics and autonomous driving where 3D object motion is a fundamental input to high-level tasks such as scene understanding, obstacle avoidance or path planning [5,7,10,28]. In this paper, we are interested in 3D scene flow estimation with a focus on autonomous driving. While a number of methods have recently demonstrated impressive performance in this context [25,35,37,39], none of them explicitly takes advantage of the fact that such scenes can often be considered as a small collection of independently moving 3D objects which include the background motion caused by the moving camera itself. See Fig. 1 for an illustration. Furthermore, due to the absence of realistic benchmarks with scene flow ground truth, quantitative evaluations are restricted to synthetic images [18, 39] or static scenes such as the ones provided by the Middlebury stereo benchmark [27] or the KITTI stereo and optical flow evaluation [12].

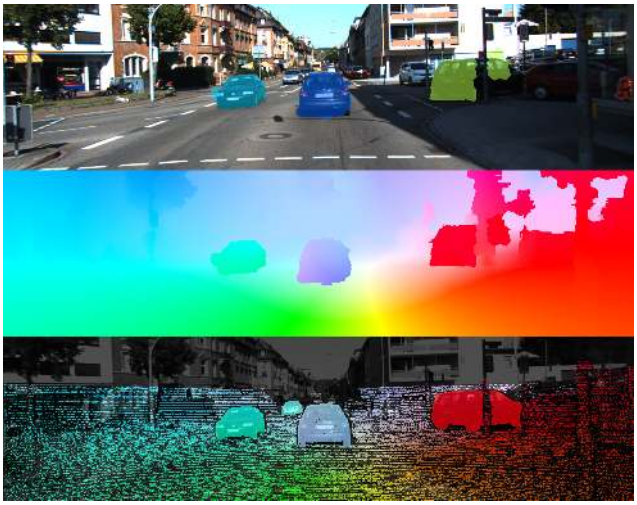

Figure 1: **Scene Flow Results on the proposed Dataset.** Top-to-bottom: Estimated moving objects with background object in transparent, flow results and flow ground truth.

The contribution of this paper is twofold: First, we propose a slanted-plane scene flow model that explicitly reasons about objects while not relying on particular shape models or pre-trained detectors. In contrast to [35, 37], we model the 3D structure of the scene as a collection of planar patches and the 3D motion of these patches by a small number of rigidly moving objects which we optimize jointly. This significantly reduces the parameter space and constrains some of the problems in textureless or ambiguous regions. Besides, our method also outputs a segmentation of the scene into independently moving objects. Second, we present a novel and realistic dataset for quantitative scene flow evaluation. Towards this goal, we collected 400 highly dynamic scenes from the KITTI raw dataset and

augmented them with semi-dense scene flow ground truth: We extract disparity maps directly from the 3D information in the laser scans and fit geometrically accurate CAD models to moving 3D point clouds to obtain optical flow ground truth. Furthermore, we provide a meaningful evaluation metric that considers depth and motion jointly.

The performance of the proposed method and the importance of its individual components are demonstrated with respect to a representative set of state-of-the-art baselines as well as in various ablation studies, leveraging the proposed dataset, the KITTI stereo and optical flow benchmark and a synthetic sequence of a rotating sphere. Besides demonstrating the value of our assumptions for this task, our experiments also show that the proposed benchmark offers novel challenges which are not handled by any existing scene flow or optical flow algorithm. We make our code[1] and dataset[2] online available.

2. Related Work

In this section, we first review related work on scene flow estimation followed by an overview over existing datasets for benchmarking scene flow approaches.

Methods: *Scene flow* is commonly defined as a flow field describing the 3D motion at every point in the scene. Following the seminal work by Vedula et al. [33, 34], the problem is traditionally formulated in a variational setting [1, 18, 23, 32, 36, 38] where optimization proceeds in a coarse-to-fine manner and local regularizers are leveraged to encourage smoothness in depth and motion. With the advent of RGB-D sensors like the Microsoft Kinect, depth information has also become available [15, 17, 24, 39]. While the Kinect sensor works well for indoor scenes, in this paper we are interested in outdoor scene flow estimation with an application to autonomous driving [9, 25] and thus focus on appearance-based methods.

Inspired by recent trends in optical flow [21, 30, 40, 41] and stereo [2, 3, 42], Vogel et al. [35, 37] proposed a slanted-plane model which assigns each pixel to an image segment and each segment to one of several rigidly moving 3D plane proposals, thus casting the task as a discrete optimization problem which can be solved using α-expansion and QPBO [26]. Impressive performance has been demonstrated in challenging street scenes[3] as well as on the KITTI stereo and optical flow benchmarks [12]. Our method (which we termed "Object Scene Flow") is related to this line of work, but goes one step further: Following Occam's razor, we take advantage of the fact that many scenes decompose into a small number of rigidly moving objects and the background. We jointly estimate *this decomposition* as well as the motion of the objects and the plane parameters of each superpixel in the image. In contrast to [35, 37] where all shape and motion proposals are fixed a-priori, we optimize the continuous variables in our model jointly with the object assignments. Besides obtaining a segmentation of the objects according to their motion, the scene flow in our model is uniquely determined by only 4 parameters per superpixel (3 for its geometry and 1 for the object index) as well as a small number of parameters for each moving object.

Datasets: Quantitative evaluation of scene flow methods suffers from a shortage of appropriate reference data. One reason for this is that no sensor exists which is capable of capturing optical flow ground truth in complex environments. Therefore, synthetic renderings of spheres [18, 32], primitives [1, 6, 36] or simple street scenes [25, 38, 39] are typically employed to measure quantitative performance. Towards more realism, recent methods [8, 14, 24] report results on the Middlebury benchmark [27] by selecting a set of rectified stereo pairs. Similarly, the more challenging KITTI benchmark [12] has been leveraged for evaluation in [35, 37]. Unfortunately, both benchmarks provide scene flow ground truth only for rigid scenes without independently moving objects. Furthermore, the scene flow evaluations based on the Middlebury stereo dataset are restricted to motions in x-direction of the image and evaluations on KITTI treat the flow and stereo problem separately.

To the best of our knowledge, there currently does not exist any realistic benchmark dataset providing dynamic objects and ground truth for the evaluation of scene flow or optical flow approaches. In this paper, we take advantage of the KITTI raw data [11] to create a realistic and challenging scene flow benchmark with independently moving objects and annotated ground truth, comprising 200 training and 200 test scenes in total. We hope that our data collection will stimulate further research on this important topic.

The remainder of this paper is structured as follows: We first introduce our "Object Scene Flow" approach in Section 3, followed by details on the ground truth annotation process for our dataset in Section 4. Finally, we show quantitative and qualitative results of our method and several state-of-the-art baselines on three datasets in Section 5. We conclude with an outlook on future work in Section 6.

3. Object Scene Flow

We focus on the classical scene flow setting where the input is given by two consecutive stereo image pairs of a calibrated camera and the task is to determine the 3D location and 3D flow of each pixel in a reference frame. We employ a slanted-plane model, i.e., we assume that the 3D structure of the scene can be approximated by a set of piecewise planar superpixels [41]. Furthermore, we assume a finite number of ridigly moving objects in the scene. It is

[1] http://www.cvlibs.net/projects/objectsceneflow
[2] http://www.cvlibs.net/datasets/kitti/
[3] http://hci.iwr.uni-heidelberg.de/Benchmarks/

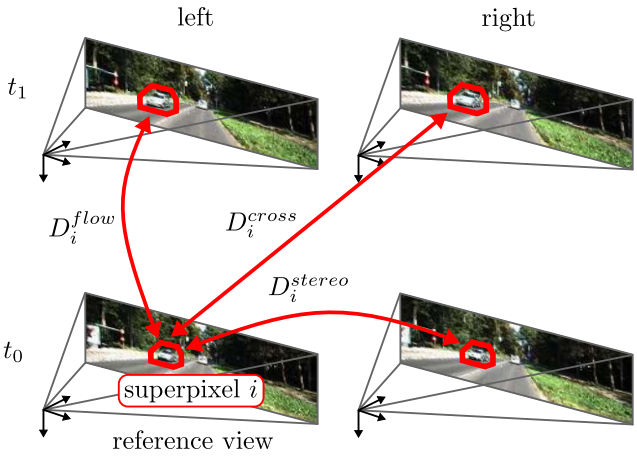

Figure 2: **Data Term.** Each superpixel i in the reference view is modeled as a rigidly moving 3D plane and warped into each other image to calculate matching costs. Each of the superpixels is associated with a 3D plane variable and a pointer to an object hypothesis comprising its rigid motion.

important to note that the static elements in the scene (the "background") can be easily handled as yet another object in our formulation as these elements move rigidly with respect to the observer.

More formally, let $\mathcal{S}$ and $\mathcal{O}$ denote the set of superpixels and objects, respectively. Each superpixel $i \in \mathcal{S}$ is associated with a region $\mathcal{R}_i$ in the image and a random variable $\mathbf{s}_i = (\mathbf{n}_i, k_i)^T$ where $\mathbf{n}_i \in \mathbb{R}^3$ describes a plane in 3D ($\mathbf{n}_i^T \mathbf{x} = 1$ for $\mathbf{x} \in \mathbb{R}^3$ on the plane) and $k_i \in \{1, \dots, |\mathcal{O}|\}$ indexes the object which the superpixel is associated with. Each object $i \in \mathcal{O}$ is associated with a random variable $\mathbf{o}_i \in SE(3)$ which describes its rigid body motion in 3D. Note that each superpixel associated with object i inherits its rigid motion parameters $\mathbf{o}_i \in SE(3)$. In combination with the plane parameters $\mathbf{n}_i$, this fully determines the 3D scene flow at each pixel inside the superpixel.

Given the left and right input images of two consecutive frames t_0 and t_1, our goal is to infer the 3D geometry of each superpixel $\mathbf{n}_i$, the association to objects k_i and the rigid body motion of each object $\mathbf{o}_i$. We specify our CRF in terms of the following energy function

$$E(\mathbf{s}, \mathbf{o}) = \sum_{i \in \mathcal{S}} \underbrace{\varphi_i(\mathbf{s}_i, \mathbf{o})}_{\text{data}} + \sum_{i \sim j} \underbrace{\psi_{ij}(\mathbf{s}_i, \mathbf{s}_j)}_{\text{smoothness}} \qquad (1)$$

where $\mathbf{s} = \{\mathbf{s}_i | i \in \mathcal{S}\}$, $\mathbf{o} = \{\mathbf{o}_i | i \in \mathcal{O}\}$, and $i \sim j$ denotes the set of adjacent superpixels in $\mathcal{S}$.

3.1. Data Term

The data term models the assumption that corresponding points across the four images should be similar in appearance. As k_i determines the association of superpixel i to an object, our data term decomposes into pairwise potentials

$$\varphi_i(\mathbf{s}_i, \mathbf{o}) = \sum_{j \in \mathcal{O}} [k_i = j] \cdot D_i(\mathbf{n}_i, \mathbf{o}_j) \qquad (2)$$

where $[\cdot]$ denotes the Iverson bracket and $D_i(\mathbf{n}, \mathbf{o})$ represents the data term at superpixel i which depends on plane parameters $\mathbf{n}$ and rigid body motion $\mathbf{o}$. The data term itself is composed of a stereo, flow and a cross term, calculated between a reference view (left image at t_0) and all other images as illustrated in Fig. 2:

$$D_i(\mathbf{n}, \mathbf{o}) = D_i^{\text{stereo}}(\mathbf{n}, \mathbf{o}) + D_i^{\text{flow}}(\mathbf{n}, \mathbf{o}) + D_i^{\text{cross}}(\mathbf{n}, \mathbf{o})$$

Each of these terms is defined by summing the matching costs of all pixels $\mathbf{p}$ inside superpixel i, where matching costs are computed by warping each pixel according to the homography induced by the associated object $\mathbf{o}$:

$$D_i^{\text{x}}(\mathbf{n}, \mathbf{o}) = \sum_{\mathbf{p} \in \mathcal{R}_i} C_{\text{x}}\big(\mathbf{p}, \underbrace{\mathbf{K}\left(\mathbf{R}_{\text{x}}(\mathbf{o}) - \mathbf{t}_{\text{x}}(\mathbf{o}) \cdot \mathbf{n}^T\right)\mathbf{K}^{-1}}_{3\times3 \text{ homography}} \mathbf{p}\big)$$

Here, $\text{x} \in \{\text{stereo}, \text{flow}, \text{cross}\}$, $\mathbf{K} \in \mathbb{R}^{3\times3}$ denotes the camera calibration matrix and $[\mathbf{R}_{\text{x}}(\mathbf{o})|\mathbf{t}_{\text{x}}(\mathbf{o})] \in \mathbb{R}^{3\times4}$ maps a 3D point in reference coordinates to a 3D point in another camera coordinate system according to the extrinsic camera calibration and the rigid motion $\mathbf{o}$. The matching cost $C_{\text{x}}(\mathbf{p}, \mathbf{q})$ returns a dissimilarity value between a pixel at location $\mathbf{p} \in \mathbb{R}^2$ in the reference image and a pixel at location $\mathbf{q} \in \mathbb{R}^2$ in the target image. In our model, we take advantage of dense as well as sparsely matched image features and define $C_{\text{x}}(\mathbf{p}, \mathbf{q})$ as a weighted sum of these two:

$$C_{\text{x}}(\mathbf{p}, \mathbf{q}) = \theta_{1,\text{x}}\, C_{\text{x}}^{\text{dense}}(\mathbf{p}, \mathbf{q}) + \theta_{2,\text{x}}\, C_{\text{x}}^{\text{sparse}}(\mathbf{p}, \mathbf{q})$$

For $C_{\text{x}}^{\text{dense}}(\mathbf{p}, \mathbf{q})$ we leverage the Hamming distance of the respective 5×5 Census descriptors [43], truncated at C_{max}. We further use an outlier value of $C = C_{\text{max}}$ for $\mathbf{q}$'s leaving the image domain. The same parameter value C_{max} is used for stereo, flow and cross terms. Our sparse matching term is defined as

$$C_{\text{x}}^{\text{sparse}}(\mathbf{p}, \mathbf{q}) = \begin{cases} \rho_{\tau_1}\left(\|\pi_{\text{x}}(\mathbf{p}) - \mathbf{q}\|_2\right) & \text{if } \mathbf{p} \in \Pi_{\text{x}} \\ 0 & \text{otherwise} \end{cases}$$

where $\pi_{\text{x}}(\mathbf{p})$ denotes the warping of pixel $\mathbf{p}$ according to the set of sparse feature correspondences, Π_{x} is the set of pixels in the reference image for which correspondences have been established, and $\rho_\tau(x)$ denotes the truncated l_1 penalty function $\rho_\tau(x) = \min(|x|, \tau)$. Details on the sparse feature correspondences we use will be given in Section 5.

3.2. Smoothness Term

The smoothness term encourages coherence of adjacent superpixels in terms of depth, orientation and motion. Our

smoothness potential $\psi_{ij}(\mathbf{s}_i, \mathbf{s}_j)$ decomposes as:

$$\psi_{ij}(\mathbf{s}_i, \mathbf{s}_j) = \theta_3 \, \psi_{ij}^{\text{depth}}(\mathbf{n}_i, \mathbf{n}_j) + \theta_4 \, \psi_{ij}^{\text{orient}}(\mathbf{n}_i, \mathbf{n}_j) + \theta_5 \, \psi_{ij}^{\text{motion}}(\mathbf{s}_i, \mathbf{s}_j) \quad (3)$$

with weights θ and

$$\psi_{ij}^{\text{depth}}(\mathbf{n}_i, \mathbf{n}_j) = \sum_{\mathbf{p} \in \mathcal{B}_{ij}} \rho_{\tau_2} \left(d(\mathbf{n}_i, \mathbf{p}) - d(\mathbf{n}_j, \mathbf{p}) \right)$$

$$\psi_{ij}^{\text{orient}}(\mathbf{n}_i, \mathbf{n}_j) = \rho_{\tau_3} \left(1 - |\mathbf{n}_i^T \mathbf{n}_j| / (\|\mathbf{n}_i\| \|\mathbf{n}_j\|) \right)$$

$$\psi_{ij}^{\text{motion}}(\mathbf{s}_i, \mathbf{s}_j) = w(\mathbf{n}_i, \mathbf{n}_j) \cdot [k_i \neq k_j]$$

Here, $d(\mathbf{n}, \mathbf{p})$ denotes the disparity of plane $\mathbf{n}$ at pixel $\mathbf{p}$ in the reference image, $\mathcal{B}_{ij}$ is the set of shared boundary pixels between superpixel i and superpixel j, and ρ is the robust l_1 penalty as defined above. The weight $w(\cdot, \cdot)$ is defined as

$$w(\mathbf{n}_i, \mathbf{n}_j) = \exp\left(-\frac{\lambda}{|\mathcal{B}_{ij}|} \sum_{\mathbf{p} \in \mathcal{B}_{ij}} \left(d(\mathbf{n}_i, \mathbf{p}) - d(\mathbf{n}_j, \mathbf{p}) \right)^2 \right) \times |\mathbf{n}_i^T \mathbf{n}_j| / (\|\mathbf{n}_i\| \|\mathbf{n}_j\|)$$

and encodes our belief that motion boundaries are more likely to occur at 3D folds or discontinuities than within smooth surfaces.

3.3. Inference

Optimization of the discrete-continuous CRF specified in Eq. 1 with respect to all superpixels and objects is an NP-hard problem and we leverage max-product particle belief propagation (MP-PBP) [22, 31] using sequential tree-reweighted message passing (TRW-S) [19] for the inner loop to find an approximate solution. We use 30 shape particles per superpixel, five objects, ten motion particles per object and 50 iterations of MP-PBP. All motion particles and half of the shape particles are drawn from a normal distribution centered at the MAP solution of the last iteration. The remaining shape particles are proposed using the plane parameters from spatially neighboring superpixels. Both strategies complement each other and we found their combination important for efficiently exploring the search space. We initialize all superpixels and their shapes using the StereoSLIC algorithm [41]. Rigid body motions are initialized by greedily extracting motion estimates from sparse scene flow vectors [13] as follows: We iteratively estimate rigid body motions using the 3-point RANSAC algorithm and find subsets with a large number of inliers using non-maxima suppression. For further details, we refer the reader to the supplementary material[4].

4. Scene Flow Dataset and Annotation

In absence of appropriate public datasets we annotated 400 dynamic scenes from the KITTI raw dataset with optical flow and disparity ground truth in two consecutive frames. The process of ground truth generation is especially challenging in the presence of individually moving objects since they cannot be easily recovered from laser scanner data alone due to the rolling shutter of the Velodyne and the low framerate (10 fps). Our annotation work-flow consists of two major steps: First, we recover the static background of the scene by removing all dynamic objects and compensating for the vehicle's egomotion. Second, we re-insert the dynamic objects by fitting detailed CAD models to the point clouds in each frame.

[4] http://www.cvlibs.net/projects/objectsceneflow

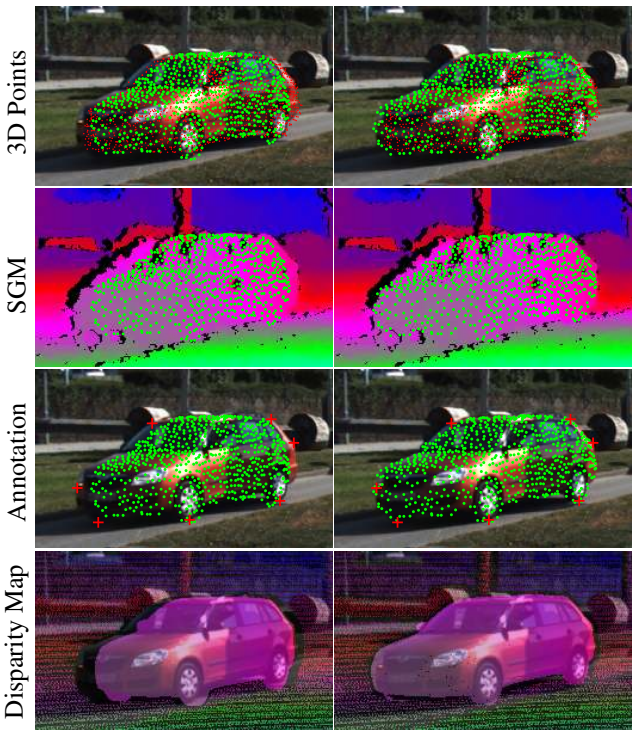

Figure 3: **Annotation.** This figure shows the subsampled CAD model (green), the observations used for registering the model (red), as well as the corresponding disparity map (last row) before (left) and after (right) minimizing Eq. 4.

4.1. Static Scene Elements

In order to derive a dense point cloud of the static scene content the laser scans are first corrected for the rolling shutter effect using the camera motion and the timestamps of the individual laser measurements. We found that neither the GPS/IMU system of the KITTI car nor ICP fitting of 3D point clouds alone yields sufficiently accurate motion estimates and thus combine both techniques using non-linear least-squares optimization to retrieve a highly accurate and consistent registration of the individual scans. Overall, we accumulate 7 scans over time in a common coordinate sys-

tem. We further remove all 3D points belonging to moving objects using the 3D bounding box annotations provided on the KITTI website[5].

4.2. Moving Objects

As the dynamic elements in the scene cannot be recovered from 3D laser measurements alone, we leverage detailed 3D CAD models from Google 3D Warehouse[6] for this purpose. It is important to note that given the limited measurement accuracy of stereo techniques our 3D CAD models are not required to be millimeter-accurate, which would be intractable considering the broad variety of vehicles in KITTI. Instead, we select the most similar model from a limited but diverse set of 16 vehicles which we illustrate in the supplementary material. For each model, we obtain a 3D point cloud by uniformly sampling $\sim 3,000$ points from all faces of the CAD model. We use this point cloud for fitting the 3D CAD model to both frames of the sequence using 2D and 3D measurements as illustrated in Fig. 3.

More specifically, for each dynamic object in the scene, we estimate a 3D similarity transformation defining the pose and scale of the 3D model in the first frame as well as the 3D rigid body motion of the object, yielding a 15-dimensional parameter vector $\xi \in \mathbb{R}^{15}$. We leverage three different types of observations: First, we accumulate 3D points belonging to a moving object over all frames using the annotated 3D bounding boxes. Second, we incorporate disparity estimates computed by semi-global matching (SGM) [16]. While SGM estimates are not always reliable, we only optimize for a very small number of parameters and found (by manual verification) that including this term as a weak prior improves results. As a third observation, we introduce manually annotated correspondences between geometrically meaningful parts of the 3D CAD model and the corresponding image coordinates in both frames. We found that including 5 to 10 such correspondences per object is sufficient for obtaining accurate optical flow ground truth.

We obtain the transformation parameters ξ by minimizing the following energy function

$$E(\xi) = \sum_{t \in \{1,2\}} E_t^{3D} + E_t^{SGM} + E_t^{2D} \qquad (4)$$

where t is the frame index and E_t are the energy terms corresponding to each of the observations, see Fig. 3 for an illustration. More specifically, E_t^{3D} denotes the average truncated l_2 distance between the 3D laser points inside the object's 3D bounding box and their nearest neighbors in the CAD model, E_t^{SGM} represents the truncated l_1 distance between the disparity map induced by the CAD model and the SGM measurements and E_t^{2D} is the quadratic $2D$ error with respect to the selected $2D-3D$ correspondences in frame t. Our optimization scheme alternates between minimizing Eq. 4 with respect to ξ using non-linear least-squares and updating all nearest neighbor associations until convergence. The weights of the terms are chosen to ensure a dominating influence of the manual input.

For generating the final disparity and optical flow maps we project a more densely sampled 3D CAD model into all four images according to the estimated ξ. For resolving intra- and inter-occlusions of objects we leverage OpenGL's z-buffer. Finally, non-rigidly moving objects like pedestrians or bicyclists and erroneous regions in the laser scans are manually masked. All resulting flow and disparity maps are validated by visual inspection. In addition, critical cases are identified and excluded by sparse, manually annotated control points. While we empirically found that for most parts our ground truth is at least 3 pixels accurate, we observed that very large motions at the image boundaries (up to 500 pixels) degrade the accuracy of the ground truth. We thus design a scene flow evaluation metric which takes these error characteristics into account as discussed below.

5. Experimental Results

This section provides a thorough quantitative and qualitative analysis of our model on the proposed scene flow dataset, the KITTI stereo/flow evaluation [12] as well as the synthetic sphere sequence of Huguet et al. [18]. As input to our method, we leverage sparse optical flow from feature matches [13] and SGM disparity maps [16] for both rectified frames. Sparse cross features are computed by combining the optical flow matches with valid disparities from the SGM maps. We obtain superpixels using StereoSLIC [41] and initialize the rigid motion parameters of all objects in the scene by greedily extracting rigid body motion hypotheses using the 3-point RANSAC algorithm implemented in [13]. In order to obtain the model parameters $\{\theta\}$ and $\{\tau\}$, we perform block coordinate descent on a subset of 30 randomly selected training images. For details on the estimated parameter values we refer the reader to the supplementary material.

Evaluation Protocol: For our results on the KITTI benchmark we follow the standard evaluation protocol and provide stereo and optical flow outliers separately using an error threshold of 3 pixels. For the proposed scene flow dataset, we annotated a total of 200 training and 200 test scenes from the KITTI raw dataset [11] using the method described in Section 4 and evaluate both disparity errors and the flow at each valid ground truth pixel in the reference view. We only count errors if the disparity or flow exceeds 3 pixels *and* 5% of its true value. Empirically, we found that this combination ensures an evaluation which is faith-

[5] http://www.cvlibs.net/datasets/kitti/raw_data.php
[6] https://3dwarehouse.sketchup.com/

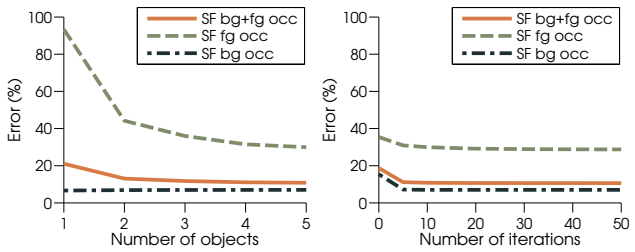

Figure 4: **Performance.** This figure shows the scene flow errors of our method on the proposed dataset with respect to the number of object proposals and MP-PBP iterations.

	[32]	[18]	[39]	[37]	Ours
RMSE 2D Flow	0.63	0.69	0.77	0.63	**0.55**
RMSE Disparity	3.8	3.8	10.9	2.84	**2.58**
RMSE Scene Flow	1.76	2.51	2.55	1.73	**0.75**

Table 1: **RMS Errors on the "Sphere" Sequence [18].**

ful with respect to the annotation errors in the ground truth. For methods which provide their second disparity estimate directly in the second frame we leverage the estimated optical flow for mapping the disparity values to the corresponding pixels in the first frame and apply background interpolation [12] for all missing pixels. Thus at each pixel in the reference view, we obtain four values which we evaluate and which uniquely determine the 3D scene flow: two disparity values (first and second frame) and the flow in u- and v-direction. We also evaluate the combination of all three measures in a single scene flow metric which considers only pixels with correct disparities *and* flow as correct. We evaluate results by averaging errors over all image regions as well as over all regions which do not leave the image domain.

Ablation Studies: We first assess the contribution of each individual term in our energy formulation in Eq. 1 on the proposed scene flow dataset. Table 2 (lower part) shows the results when evaluating the whole image. Results for non-occluded regions can be found in the supplementary material. The columns show errors in terms of disparity ("D1", "D2"), flow ("Fl") and scene flow ("SF") using the conventions specified above. For each modality we provide results in terms of the background ("bg"), foreground ("fg") as well as the combination of both ("bg+fg"). The first row of the table shows the results of our model when only including the unary terms. The remaining rows show results for different combinations of unary and pairwise terms with the full model at the bottom. While Census proves to be a stronger feature than sparse optical flow in combination with SGM, their combination outperforms each of them individually in terms of scene flow error. The experiments also reveal that the boundary term is the strongest pairwise cue while again the combination of all pairwise terms yields the overall best scene flow results.

Next, we investigate the performance of our full model with respect to the size of the object set in Fig. 4 (left). Towards this goal, we decrease the number of allowed object hypotheses in our model from 5 to 1. This plot affirms our assumption that the outdoor scenes we consider can be well described by a small number of rigidly moving objects. Finally, Fig. 4 (right) shows the performance of our method with respect to the number of MP-PBP iterations. While we use 50 iterations for all our experiments in practice, this plot shows that 10 iterations are sufficient for achieving almost optimal performance under our model.

Baseline Results: Table 2 (upper part) compares the results of our method (last line) to several baselines on our novel scene flow dataset. We interpolate the results of sparse and semi-dense methods using the KITTI stereo/flow development kit to ensure a fair comparison. Besides the classic variational approach of Huguet et al. [18], we also compute results for the sparse scene flow method of Cech et al. [6]. We further construct several baselines by combining two state-of-the-art optical flow algorithms [4, 29] with disparity estimates in both frames obtained using semi-global matching (SGM) [16] which also serves as input to our method. As a representative for RGB-D based algorithms we show the results of Hornacek's Sphere Flow [17] which have been provided to us by the authors. To emulate the required depth component we reproject all valid pixels of SGM disparity maps into 3D. Finally, we also include the results of Vogel's piece-wise rigid scene flow (PRSF) approach [37]. We note that the proposed approach strictly outperforms all baselines with PRSF being the closest competitor.

Qualitative Results: Fig. 5 provides qualitative results for some of the scenes in the proposed dataset using similar color mappings as in KITTI [12]. Note however, that the percentage error is mapped so that inliers according to our scene flow metric are depicted in blue shades. As evidenced by the error images, our method is able to recover the correct disparity and flow in a variety of challenging scenes. Even objects which are not perfectly rigid (and for which no ground truth exists) such as the bicyclist in subfigure (2,2) are robustly detected by our method. Some failure cases of our method are illustrated below the horizontal line. Those scenes are extremely challenging due to difficult lighting conditions or quickly moving objects in the vicinity of the observer and none of the algorithms in our evaluation was able to cope with these challenges.

Results on the KITTI Benchmark: We also evaluated our model on the static scenes of the KITTI stereo and optical flow benchmark [12] and rank amongst the top 5 methods in each category. Using the 3 pixels evaluation thresh-

	D1			D2			Fl			SF		
	bg	*fg*	*bg+fg*	*bg*	*fg*	*bg+fg*	*bg*	*fg*	*bg+fg*	*bg*	*fg*	*bg+fg*
Huguet [18]	27.31	21.71	26.38	59.51	44.92	57.08	50.06	47.57	49.64	67.69	64.03	67.08
GCSF [6]	11.64	27.11	14.21	32.94	35.76	33.41	47.38	45.07	47.00	52.92	59.11	53.95
SGM [16] + LDOF [4]	5.15	15.27	6.83	29.58	23.47	28.56	41.07	35.52	40.15	43.99	44.77	44.12
SGM [16] + Sun [29]	5.15	15.27	6.83	28.77	25.64	28.25	34.83	45.46	36.60	38.21	53.03	40.68
SGM [16] + Sphere Flow [17]	5.15	15.27	6.83	14.10	23.12	15.60	20.91	28.89	22.24	23.09	37.11	25.42
PRSF [37]	4.74	13.73	6.24	11.14	20.47	12.69	11.73	27.72	14.39	13.49	33.71	16.85
Unary (SGM+SpF)	5.26	14.40	6.78	6.48	28.67	10.17	8.25	37.03	13.04	10.15	42.58	15.54
Unary (Census)	6.89	19.22	8.95	8.01	25.35	10.90	7.71	26.63	10.86	9.69	35.34	13.96
Unary (All)	6.08	16.70	7.85	7.15	23.72	9.91	7.17	25.97	10.29	8.93	33.79	13.06
Unary (All) + Pair (Boundary)	4.58	**10.60**	**5.58**	5.58	**19.34**	7.87	5.88	23.90	8.88	7.28	30.00	11.06
Unary (All) + Pair (Normal)	5.70	16.01	7.42	6.77	23.42	9.54	6.86	25.68	9.99	8.54	33.47	12.69
Unary (All) + Pair (Object)	6.23	17.86	8.16	7.24	23.99	10.03	7.14	24.92	10.10	8.93	33.09	12.95
Unary (SGM+SpF) + Pair (All)	6.60	25.71	9.78	7.81	33.94	12.16	9.54	37.43	14.19	11.59	44.25	17.03
Unary (Census) + Pair (All)	4.67	12.37	5.95	5.58	19.74	7.94	5.66	22.57	8.47	7.09	29.37	10.79
Unary (All) + Pair (All) Fast	**4.45**	12.73	5.83	**5.41**	20.12	7.85	5.71	23.57	8.68	7.14	30.48	11.03
Unary (All) + Pair (All)	4.54	12.03	5.79	5.45	19.41	**7.77**	**5.62**	**22.18**	**8.37**	**7.01**	**28.76**	**10.63**

Table 2: **Quantitative Results on the Proposed Scene Flow Dataset.** This table shows the disparity (D1/D2), flow (Fl) and scene flow (SF) errors averaged over all 200 test images. For each modality we separately provide the errors for the background region (bg), all foreground objects (fg) as well as all pixels in the image (bg+fg).

old, we achieve 3.28 % errors for stereo and 3.47 % errors for optical flow, comparing favorably with respect to PRSF. All details and the full result tables are provided in the supplementary material.

Results on the "Sphere" Dataset: For completeness we provide results of our method on the synthetic "Sphere" dataset by Huguet et al. [18]. As this dataset resembles a random dot stereogram, appearance does not convey information about object boundaries. We thus modify the StereoSLIC algorithm to consider dense optical flow instead of disparity and provide the Horn-Schunck results of Sun et al. [29] as input. As illustrated in Table 1 our method performs surprisingly well despite the fact that we restrict the scene to only 200 planar superpixels. Qualitative results and error maps are shown in the supplementary material.

Runtime: Our non-optimized MATLAB implementation with C++ wrappers requires on average 60 seconds for each of the 50 MP-PBP iterations. This yields a total runtime of 50 minutes for processing one scene (4 images) on a single i7 core running at 3.0 Ghz. By restricting the number of shape and motion particles to 10 and 5, respectively, and by setting the number of MP-PBP iterations to 10, we are able to reduce the total runtime of our algorithm to 120 seconds. The entry 'Fast' in Table 2 shows that the modified version performs only slightly worse than the full method presented in the paper, but still compares favorably with respect to the closest state-of-the-art competitor.

Supplementary Material: We encourage the reader to have a look at the supplementary material[7] which provides an analysis of performance with respect to variation of parameters, and additional quantitative and qualitative results.

6. Conclusion

We have demonstrated the benefits of modeling dynamic outdoor scenes as a collection of rigid objects. By reasoning jointly about this decomposition as well as the geometry and motion of a small number of objects in the scene the proposed model is able to produce accurate dense 3D scene flow estimates, comparing favorably with current state-of-the-art on several datasets. We have further introduced the first realistic and large-scale scene flow dataset with ground truth annotations for all static and dynamic objects in the scene. Compared to KITTI stereo/flow, our benchmark provides dynamic objects and a dedicated scene flow measure as well as novel challenges to the community. In particular, we found that none of the existing optical flow or scene flow methods is able to cope with the extreme motions produced by moving objects in some of our scenes. A second source of failure are textureless and reflecting surfaces where often both, stereo matching and optical flow estimation fails. We conjecture that more expressive priors are required to overcome these challenges and believe that our dataset will stimulate further research towards solving these problems.

[7] http://www.cvlibs.net/projects/objectsceneflow

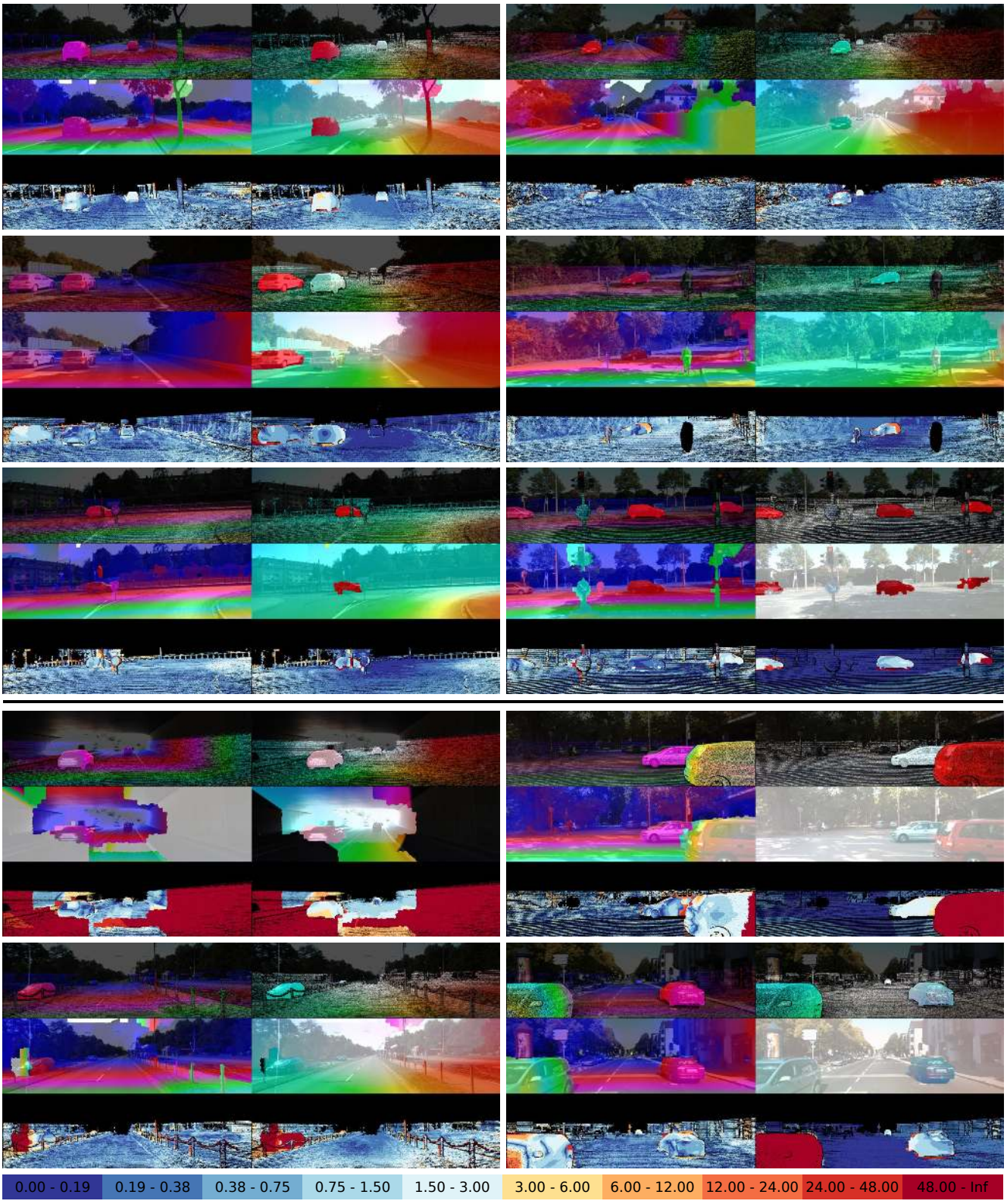

Figure 5: **Qualitiative Results.** Each subfigure shows from top-to-bottom: disparity and optical flow ground truth in the reference view, the disparity map (D1) and optical flow map (Fl) estimated by our algorithm, and the respective error images using the color scheme depicted in the legend. The four scenes below the horizontal line are failure cases. See text for details.

References

[1] T. Basha, Y. Moses, and N. Kiryati. Multi-view scene flow estimation: A view centered variational approach. *International Journal of Computer Vision (IJCV)*, 101(1):6–21, 2013. 2

[2] M. Bleyer, C. Rhemann, and C. Rother. Extracting 3D scene-consistent object proposals and depth from stereo images. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2012. 2

[3] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha. Object stereo - joint stereo matching and object segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2

[4] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 33:500–513, March 2011. 6, 7

[5] M. Buehler, K. Iagnemma, and S. Singh. The DARPA urban challenge. *DARPA Challenge*, 56, 2009. 1

[6] J. Cech, J. Sanchez-Riera, and R. P. Horaud. Scene flow estimation by growing correspondence seeds. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2, 6, 7

[7] A. Ess, B. Leibe, K. Schindler, and L. V. Gool. Robust multi-person tracking from a mobile platform. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 31:1831–1846, 2009. 1

[8] D. Ferstl, G. Riegler, M. Rther, and H. Bischof. Cp-census: A novel model for dense variational scene flow from RGB-D data. In *Proc. of the British Machine Vision Conf. (BMVC)*, 2014. 2

[9] U. Franke, C. Rabe, H. Badino, and S. Gehrig. 6D-Vision: fusion of stereo and motion for robust environment perception. In *Proc. of the DAGM Symposium on Pattern Recognition (DAGM)*, 2005. 2

[10] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3D traffic scene understanding from movable platforms. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 36(5):1012–1025, May 2014. 1

[11] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research (IJRR)*, 32(11):1231–1237, September 2013. 2, 5

[12] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 2, 5, 6

[13] A. Geiger, J. Ziegler, and C. Stiller. StereoScan: Dense 3D reconstruction in real-time. In *Proc. IEEE Intelligent Vehicles Symposium (IV)*, 2011. 4, 5

[14] S. Hadfield and R. Bowden. Scene flow estimation using intelligent cost functions. In *Proc. of the British Machine Vision Conf. (BMVC)*, 2014. 2

[15] E. Herbst, X. Ren, and D. Fox. RGB-D flow: Dense 3D motion estimation using color and depth. In *Proc. IEEE International Conf. on Robotics and Automation (ICRA)*, 2013. 2

[16] H. Hirschmueller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 30(2):328–341, 2008. 5, 6, 7

[17] M. Hornacek, A. Fitzgibbon, and C. Rother. SphereFlow: 6 DoF scene flow from RGB-D pairs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 6, 7

[18] F. Huguet and F. Devernay. A variational method for scene flow estimation from stereo sequences. In *Proc. IEEE International Conf. on Computer Vision (ICCV)*, 2007. 1, 2, 5, 6, 7

[19] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 28(10):1568–1583, 2006. 4

[20] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Company, 1983. 1

[21] T. Nir, A. M. Bruckstein, and R. Kimmel. Over-parameterized variational optical flow. *International Journal of Computer Vision (IJCV)*, 76(2):205–216, 2008. 2

[22] J. Pacheco, S. Zuffi, M. J. Black, and E. Sudderth. Preserving modes and messages via diverse particle selection. In *Proc. of the International Conf. on Machine learning (ICML)*, 2014. 4

[23] J.-P. Pons, R. Keriven, and O. Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *International Journal of Computer Vision (IJCV)*, 72(2):179–193, 2007. 2

[24] J. Quiroga, T. Brox, F. Devernay, and J. L. Crowley. Dense semi-rigid scene flow estimation from RGB-D images. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 2

[25] C. Rabe, T. Mueller, A. Wedel, and U. Franke. Dense, robust, and accurate motion field estimation from stereo image sequences in real-time. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2010. 1, 2

[26] C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer. Optimizing binary mrfs via extended roof duality. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007. 2

[27] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision (IJCV)*, 47:7–42, 2002. 1, 2

[28] A. Shashua, Y. Gdalyahu, and G. Hayun. Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. In *Proc. IEEE Intelligent Vehicles Symposium (IV)*, 2004. 1

[29] D. Sun, S. Roth, and M. J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision (IJCV)*, 106(2):115–137, 2013. 6, 7

[30] D. Sun, J. Wulff, E. Sudderth, H. Pfister, and M. Black. A fully-connected layered model of foreground and background flow. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[31] H. Trinh and D. McAllester. Unsupervised learning of stereo vision with monocular cues. In *Proc. of the British Machine Vision Conf. (BMVC)*, 2009. 4

[32] L. Valgaerts, A. Bruhn, H. Zimmer, J. Weickert, C. Stoll, and C. Theobalt. Joint estimation of motion, structure and geometry from stereo sequences. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2010. 2, 6

[33] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1999. 2

[34] S. Vedula, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 27(3):475–480, 2005. 2

[35] C. Vogel, S. Roth, and K. Schindler. View-consistent 3D scene flow estimation over multiple frames. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 1, 2

[36] C. Vogel, K. Schindler, and S. Roth. 3D scene flow estimation with a rigid motion prior. In *Proc. IEEE International Conf. on Computer Vision (ICCV)*, 2011. 2

[37] C. Vogel, K. Schindler, and S. Roth. Piecewise rigid scene flow. In *Proc. IEEE International Conf. on Computer Vision (ICCV)*, 2013. 1, 2, 6, 7

[38] A. Wedel, T. Brox, T. Vaudrey, C. Rabe, U. Franke, and D. Cremers. Stereoscopic scene flow computation for 3D motion understanding. *International Journal of Computer Vision (IJCV)*, 95(1):29–51, 2011. 2

[39] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, U. Franke, and D. Cremers. Efficient dense scene flow from sparse or dense stereo data. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2008. 1, 2, 6

[40] J. Wulff and M. J. Black. Modeling blurred video with layers. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 2

[41] K. Yamaguchi, D. McAllester, and R. Urtasun. Robust monocular epipolar flow estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2, 4, 5

[42] K. Yamaguchi, D. McAllester, and R. Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 2

[43] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 1994. 3