Object Scene Flow

Moritz Menze[a,*], Christian Heipke[a], Andreas Geiger[b,c]

[a] *Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Nienburger Str. 1, D-30167 Hannover, Germany*
[b] *Autonomous Vision Group, Max Planck Institute for Intelligent Systems, Spemannstr. 41, D-72076 Tübingen, Germany*
[c] *Computer Vision and Geometry Lab, ETH Zürich, Universitätstrasse 6, CH-8092, Zürich, Switzerland*

Abstract

This work investigates the estimation of dense three-dimensional motion fields, commonly referred to as *scene flow*. While great progress has been made in recent years, large displacements and adverse imaging conditions as observed in natural outdoor environments are still very challenging for current approaches to reconstruction and motion estimation. In this paper, we propose a unified random field model which reasons jointly about 3D scene flow as well as the location, shape and motion of vehicles in the observed scene. We formulate the problem as the task of decomposing the scene into a small number of rigidly moving objects sharing the same motion parameters. Thus, our formulation effectively introduces long-range spatial dependencies which commonly employed local rigidity priors are lacking. Our inference algorithm then estimates the association of image segments and object hypotheses together with their three-dimensional shape and motion. We demonstrate the potential of the proposed approach by introducing a novel challenging scene flow benchmark which allows for a thorough comparison of the proposed scene flow approach with respect to various baseline models. In contrast to previous benchmarks, our evaluation is the first to provide stereo and optical flow ground truth for dynamic real-world urban scenes at large scale. Our experiments reveal that rigid motion segmentation can be utilized as an effective regularizer for the scene flow problem, improving upon existing two-frame scene flow methods. At the same time, our method yields plausible object segmentations without

*Corresponding author
Email address: moritz.menze@alumni-uni-hannover.de (Moritz Menze)

requiring an explicitly trained recognition model for a specific object class.

Keywords: Scene Flow, Motion Estimation, Motion Segmentation, 3D Reconstruction, Active Shape Model, Object Detection

1. Introduction

Scene flow estimation provides valuable information about the dynamic nature of our three-dimensional environment. In particular, the three-dimensional scene flow field comprises all 3D motion vectors of a densely reconstructed 3D surface model, which is moving with respect to the camera. Recovering scene flow from image observations, however, is an inherently ill-posed inverse problem, requiring the development of appropriate priors for regularizing the space of solutions.

In addition to the inherent academic interest in perceiving systems, image-based scene flow estimation is relevant for a broad range of applications. While active sensors are a strong competitor in many fields, image sequences contain valuable dynamic information. Automatic navigation of autonomous platforms (Geiger et al., 2014; Zhang et al., 2013) is just one example requiring a detailed dynamic perception of the 3D environment. While warning and avoidance of moving obstacles is already part of advanced driver assistance systems, existing solutions are still restricted to certain types of objects, limited speed and ranges. The safe interaction of robots with their environment also requires up-to-date and precise information about their surroundings. Furthermore, motion cues are important for action and activity recognition for example in video surveillance applications. All of these tasks benefit from an improved perception of surrounding shapes and motions.

In this work, we propose a consistent model allowing for joint inference of both entities. In particular, we propose a unified random field model which reasons jointly about 3D scene flow as well as the location, shape and motion of vehicles in the observed scene. We formulate the problem as the task of decomposing the scene into a small number of rigidly moving objects sharing the same motion parameters. Our inference algorithm estimates the association of image segments and object hypotheses together with their three-dimensional shape and motion. We extend our model to

jointly estimate the parametrized 3D shape of each vehicle in the scene. To evaluate our approach, we develop a comprehensive dataset and evaluation, the KITTI 2015 scene flow benchmark[1], allowing for detailed quantitative analysis of the results and an in-depth comparison to the state-of-the-art.

1.1. Related Work

Image-based methods for scene flow estimation can be categorized into variational and discrete optimization approaches. With the advent of consumer grade active sensors like the Microsoft Kinect, depth information has become readily available and is leveraged by a number of recent scene flow approaches, e.g., Herbst et al. (2013); Hornacek et al. (2014); Quiroga et al. (2014). While active sensors work well for indoor scenes of limited extent, the focus of this paper is on the outdoor scenario with applications to autonomous driving. Therefore, we concentrate on appearance based methods in our literature review.

1.1.1. Scene Flow Estimation

Following the seminal approaches to optical flow (Horn & Schunck, 1981) and scene flow estimation (Vedula et al., 1999, 2005), the problem of estimating a three-dimensional displacement field is traditionally formulated in a variational setting. As depth information is needed, different ways to incorporate dense reconstruction into the variational framework have been proposed. Analogous to the 2D case, optimization has to proceed in a coarse-to-fine manner to avoid local minima of the energy functional and capture large displacements.

Pons et al. (2007) alternatingly optimize the reconstruction of a surface model and the motion field. The key contribution addresses the data term. To circumvent common assumptions of similarity measures the authors propose a global prediction error evaluating the consistency of all input images, which are warped according to the reconstructed surfaces and estimated motion. The resulting algorithm appropriately handles projective distortion and partial occlusions. To regularize the results simple

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

smoothness constraints are imposed. The resulting energy functional is optimized in a coarse-to-fine gradient descent framework.

Huguet & Devernay (2007) generalize the variational optical flow method of Brox et al. (2004) to jointly infer geometry and motion. To this end, they propose a minimal representation of scene flow by four variables in the image domain. In particular, they compute the disparity at the first time step t_0, the optical flow with respect to the reference image and the disparity at the second time step t_1. Given a calibrated stereo camera, the three-dimensional scene flow can directly be computed from this representation. To jointly optimize stereo disparity and optical flow, Huguet & Devernay (2007) extend the respective data and smoothness terms to cover all sought entities and combine them in a unified energy functional. This formulation leads to four partial differential equations, which are optimized using the numerical scheme proposed by Brox et al. (2004) for 2D optical flow. Since stereo image matching typically has to deal with large displacements, a dedicated initialization procedure is required. To this end, pre-computed disparity and optical flow maps are employed. We leverage a similar strategy for the initialization of the proposed approach (see Section 2.3).

An important aspect of scene flow estimation is regularization. Basha et al. (2013) argue that smooth 3D motion fields can project to discontinuous 2D flow fields and thus propose a 3D model representing the scene as a point cloud with spatial motion vectors. This formulation allows to apply regularization directly to the three-dimensional motion vectors and to easily extend the method to a multi-view set-up. Vogel et al. (2011) replace the global total variation regularization by a piecewise rigid prior. Thus, sharp discontinuities in the scene flow field can be preserved more faithfully.

Valgaerts et al. (2010) discard the common assumption of a fully calibrated stereo rig and explicitly estimate the relative orientation between the stereo heads. Consequently, the results are only retrieved up to an unknown scale factor. The energy functional becomes more complex as it now comprises general stereo terms based on the unknown fundamental matrix and is minimized in a coarse-to-fine optimization scheme. Furthermore, the authors decouple regularization of shape and motion, as they do not assume respective discontinuities to coincide.

For reasons of computational efficiency, Wedel et al. (2008) completely decouple

shape and motion estimation and focus on the computation of the displacement vector field, which significantly increases frame rate but also discards valuable mutual constraints between both entities. Rabe et al. (2010) parallelize the required computations on a GPU. They apply Kalman filtering to each pixel individually to smooth the resulting motion vectors over longer image sequences. Aiming at high frame rates, the recent prediction-correction approach of Derome et al. (2016) follows a similar strategy. Depth maps from stereo image matching are combined with visual odometry to predict optical flow vectors. While static parts of the scene can be recovered directly, a correction step has to be applied to account for individually moving objects.

For stereo matching and optical flow estimation, several publications have demonstrated the usefulness of slanted-plane models (Bleyer et al., 2011; Yamaguchi et al., 2013). Either the visible surface of the scene or its projected motion are assumed to vary smoothly within small regions in the reference image. Extending this idea to 3D forms the basis for the currently most successful scene flow models.

Yamaguchi et al. (2014) propose a semi-dense method, which builds on the well-known semiglobal matching described in (Hirschmüller, 2008). The stereo approach is extended to incorporate a third image from the reference camera, taken at a second time step. Based on the assumption of a static scene, this additional information increases the robustness of image matching. In addition to the disparity map in the reference image, it yields an estimate of the optical flow. To smooth and extrapolate the matching results, a slanted-plane model is optimized yielding an over-segmentation of the reference view together with dense estimates of disparity and optical flow. The combined approach is referred to as slanted plane smoothing of stereo and flow (SPS-StFl). As in previous work (Yamaguchi et al., 2013) there is a purely stereoscopic variant of the approach (SPS-St) and a dedicated version which is tailored towards optical flow estimation (SPS-Fl). In a related work, Lv et al. (2016) proposed a purely continuous factor-graph optimization using a piecewise-planar scene flow model.

Vogel et al. (2013b) propose a scene flow approach assuming piece-wise rigid surfaces (PRSF). Their formulation decomposes the 3D scene into planar regions, each undergoing a rigid motion. The reference image is decomposed into segments and for each of the segments, a parametrized representation of shape and motion is retrieved.

Consequently, the number of unknowns is reduced compared to a pixel-wise representation. The smoothness assumption within each segment further implements a strong regularization. Inference in this model assigns each pixel to an image segment and each segment to one of several rigidly moving plane proposals in three-dimensional space, thus casting the task as a discrete labeling problem.

To initialize the plane proposals, 3D plane parameters and rigid body transformations are robustly fit to initial disparity and flow maps. These observations are evaluated with respect to an initial segmentation of the reference frame. During inference, the resulting planes are proposed for superpixels in the vicinity of the reference segment. An estimate of the ego-motion is introduced as another proposal.

The objective function optimized during inference is specified as a discrete conditional random field. Based on the plane proposals, the energy is approximately minimized via α-expansion and quadratic pseudo-boolean optimization (QPBO) (Rother et al., 2007). First, the association of image segments to plane proposals is found and next the assignment of pixels to image segments is refined based on the initial result. Impressive performance has been demonstrated on challenging street scenes as well as on the KITTI stereo and optical flow benchmarks (Geiger et al., 2012). In consecutive work (Vogel et al., 2014, 2015) the model has been extended to longer image sequences beyond the classical set-up of two subsequent stereo pairs.

The proposed scene flow method described in Section 2 is related to this line of work. In contrast to existing works, however, the proposed model takes advantage of the fact that many real-world scenes can be decomposed into a *small* number of rigidly moving objects including the background. In the spirit of energy-based model selection algorithms (Isack & Boykov, 2012), our parametrization allows for implicit model selection to determine the number of objects in the scene. The presented approach jointly estimates this decomposition as well as the motion of the objects and the plane parameters of each superpixel in the image. In contrast to Vogel et al. (2013b, 2014), 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 four parameters per superpixel (3 for its geometry and 1 for the object

assignment), together with a small number of parameters for each moving object. This implicitly provides a strong regularizer with respect to the types of expected scenes and introduces long-range spatial interactions into the model (i.e., distant superpixels which are assigned to the same object act upon the same rigid motion parameters). Experiments, presented in Section 4, reveal that our model yields faithful reconstructions and is able to overcome motion ambiguities, which are hard to handle with local regularizers alone.

1.1.2. Object Models

As any ill-posed problem, scene flow estimation requires appropriate regularizers to overcome ambiguities. Adequate mathematical models impose reasonably general assumptions on the observed displacement field (e.g., smoothness of surfaces). Considering specific applications such as autonomous driving, it seems promising to incorporate additional task-specific sources of information. Three-dimensional geometric object models have a long history in supporting reconstruction from images in a broad range of applications.

Pioneering work, for example by Braun et al. (1995) and Debevec et al. (1996), made use of shape primitives to support photogrammetric modeling of buildings. While modeling generic objects, like buildings, is a very challenging task by itself, there are tractable approaches to formalize the geometry of objects with moderate intra-class variability. Faces and cars are prominent examples of well-defined geometry, which are frequently addressed in the literature. A widely used representation of such geometric objects is the Active Shape Model (ASM) proposed by Cootes et al. (1995). This model is based on manually annotated, corresponding landmark points. Mean positions of these landmarks are computed from a set of annotated training examples. Principal component analysis of the training data yields characteristic deformations between similar shapes. Deformed versions of the model are computed as linear combinations of the mean shape and a weighted sum of the deformations. Thus, the model is flexible but it can only deform in accordance with the variability contained in the training data. One exemplary line of work that points out the importance of a feedback-loop between early vision and high-level interpretation was published in (Leibe et al., 2006; Thomas

et al., 2007). Based on an implicit shape model the approaches are able to transfer meta information from training images to unseen object instances. This high-level object knowledge can be employed as prior information for early vision tasks like depth reconstruction.

Recently, the integration of object models into reconstruction algorithms has attracted renewed attention. Bao et al. (2013) support a multi-view stereo approach with an object model. The authors compute a mean shape from laser scans of different instances of an object class along with a set of discrete anchor points. An object detector is applied to the input images to instantiate the model. Using HOG features, the mean shape is adapted to a newly observed instance of the object by registering the anchor points. Dame et al. (2013) also use an object detector to infer the initial pose and shape parameters of an object model, which they then optimize in a variational SLAM framework. Güney & Geiger (2015) introduce CAD shapes to support binocular stereo matching. They make use of a semantic segmentation of the reference frame to initialize and constrain object hypotheses. So-called *displets*, object-specific disparity patches, are randomly sampled from a large set of CAD models and integrated into the estimated disparity map to fill in uncertain regions. Zhou et al. (2015) optimize for the geometry across several instances of an object class. Generic object detectors in three-dimensional space are employed to bootstrap the process. As opposed to these methods, our model does not require an object detector but uses a simple, motion-based segmentation of the scene to initialize object hypotheses.

Recently, Prisacariu et al. (2013) proposed an efficient way to compress prior information from CAD models with complex shape variations using Gaussian Process Latent Variable Models. Zia et al. (2013, 2015) revisited the idea of the ASM and applied it to a set of manually annotated CAD models to derive detailed 3D geometric object class representations. While they tackle the problem of object recognition and pose estimation from single images, in this work, such models are used in the context of 3D scene flow estimation.

1.2. Contributions

A preliminary version of this article was published in (Menze & Geiger, 2015; Menze et al., 2015b). This work provides the following additional contributions: First, we combine our scene flow model with robust discrete optical flow estimates (Menze et al., 2015a), tackling the large displacement problem and challenging imaging conditions in realistic outdoor scenes. As demonstrated by our experiments, this results in reduced error rates and run times compared to the original version of the algorithm. Second, we introduce an additional term that allows to jointly infer dense three-dimensional scene flow and a parameterized reconstruction of objects. Towards this goal, we extend the representation of objects by the shape and pose parameters of an Active Shape Model of cars. Third, we provide a more detailed quantitative and qualitative comparison to competing approaches on the proposed KITTI 2015 scene flow benchmark, highlighting the benefits and drawbacks of our method. Finally, we also provide a detailed description of the newly introduced KITTI 2015 scene flow dataset and its construction, emphasizing additional features and challenges compared to earlier versions of the benchmark.

2. Method

In this work, joint estimation of three-dimensional geometry and motion of an observed scene are enabled by processing stereoscopic image sequences. We make the following general assumptions about the available input data: The relative pose of the two cameras, which are mounted rigidly with respect to each other onto a stereo rig, is assumed to be known. Based on this information the images are rectified so that epipolar lines are projected to corresponding image rows and stereo matching reduces to one-dimensional disparity estimation. Besides, the synchronization of the stereo cameras is regarded as sufficiently accurate to neglect influences induced by offset exposure.

Following prior work, we employ a slanted-plane model to capture geometry and motion. More specifically, we assume that the variable three-dimensional structure of the scene can be approximated by a set of piecewise planar surface elements, each

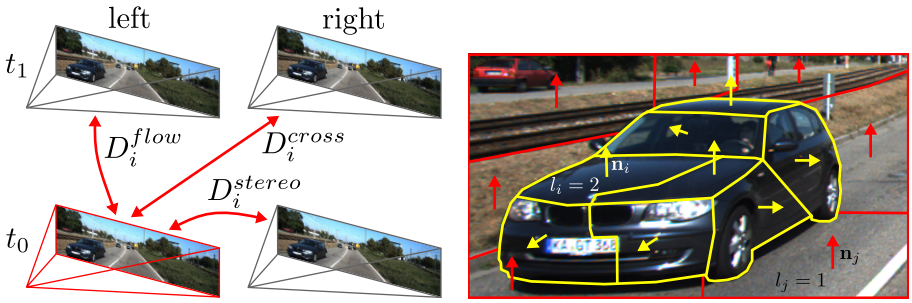
undergoing a rigid body transformation. These surface elements are associated with image segments, which completely cover the image domain of the reference view, see Figure 1 for an illustration. To capture all significant discontinuities in the sought entities, an oversegmentation of the image is carried out.

2.1. Scene Flow Model

The major novelty of the presented model is the assumption that the observed scene decomposes into a small number of *rigidly moving objects*. This is reasonable for typical traffic scenes as observed by autonomous vehicles. While pedestrians move in a non-rigid manner they are usually depicted at a scale that allows for the estimation of the dominant rigid body transformation. Surrounding vehicles fully correspond to this simple motion model. To emphasize this feature the proposed method is referred to as *Object Scene Flow (OSF)*. It is important to note that the static elements in the scene, which will be referred to as the *background*, can be easily handled as one of the objects in the proposed formulation. Like individually moving foreground objects, these parts of the scene (by definition) move rigidly with respect to the observer. For moving cameras in static environments, the background object is able to capture the complete observed motion. Based on the decomposition into objects, motion estimation simplifies to the optimization of a small number of shape and rigid body motion parameters.

To formalize our model, let us assume a superpixelization of the reference image as illustrated in Figure 1. Let $\mathcal{S}$ denote the set of superpixels and $\mathcal{O}$ denote the set of objects. Each individual superpixel $i \in \mathcal{S}$ is associated with a region $\mathcal{R}_i$ in the reference image and a random variable $\mathbf{s}_i = (\mathbf{n}_i, l_i)$. In particular, $\mathbf{n}_i \in \mathbb{R}^3$ describes a plane in 3D by its normal, scaled by the distance from the origin. Thus, $\mathbf{n}_i^T \mathbf{X} = 1$ for points $\mathbf{X} \in \mathbb{R}^3$ on the plane. The discrete label $l_i \in \{1, \ldots, |\mathcal{O}|\}$ assigns each superpixel to one of the objects. Label $l = 1$ is reserved for the static background.

Each object $k \in \mathcal{O}$ is associated with a random variable $\mathbf{o}_k = (\mathbf{R}_k, \mathbf{t}_k) \in SE(3)$ that contains a rotation matrix and a translation vector describing its rigid body motion in 3D. Each superpixel associated with object $\mathbf{o}_k$, i.e., for which $l_i = k$, inherits the rigid motion parameters from this object. In combination with the plane normal $\mathbf{n}_i$, this fully



(a) Reference View and Data Terms (b) Superpixels in the Reference View $\{(\mathbf{n}_i, l_i)\}_{i=1}^{|\mathcal{S}|}$

Figure 1: **Illustration of Data Terms and Image Segmentation.** (a) Our data term measures photo consistency between the reference view (red) and the three other views (gray). (b) Each superpixel in the reference view is represented by a plane in 3D space ($\mathbf{n}_i$) and a label l_i which determines the associated object. Here, the red superpixels have been associated with the background model ($l = 1$) and the yellow superpixels with the first object hypothesis ($l = 2$).

determines the three-dimensional scene flow of each surface element in the reference view.

The parametrization of our model is illustrated in Figure 1. Panel (a) provides an overview of the employed data terms, which are computed with respect to the reference view. Panel (b) schematically depicts the approximation of the visible surfaces with image segments. The arrows represent the estimated normals of planar segments in three-dimensional space. In the figure, all superpixels drawn in red are assigned to the background while those in yellow constitute an individually moving foreground object.

We specify our scene flow model as a conditional random field expressed via the Gibbs energy function

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

where $\mathbf{s} = \{\mathbf{s}_i \mid 1 \leq i \leq |\mathcal{S}|\}$ and $\mathbf{o} = \{\mathbf{o}_k \mid 1 \leq k \leq |\mathcal{O}|\}$. $\mathcal{N}$ denotes the set of adjacent superpixels in $\mathcal{S}$. As the neighborhood relations between image segments depend on the segmentation result, the number of neighboring superpixels differs from superpixel to superpixel and corresponds to the number of image segments sharing boundary pixels. The respective adjacency matrix is computed based on the segmentation of the

reference frame. A relative weight that trades off the influence of the data and the smoothness term is included in the individual components as described in the following sections. Given the objective function (1), our goal is to jointly infer the geometry of each segment $\mathbf{n}_i$, the association l_i of superpixels to objects, and the rigid body motion $(\mathbf{R}_k, \mathbf{t}_k)$ of each object $\mathbf{o}_k$.

To constrain the problem we make use of different observations in the data term, all of which will be explained in detail in the next section. The smoothness term will be described in Section 2.1.2 and implements the model assumptions that depth and motion vary smoothly between neighboring image segments except for the case of abrupt, significant changes of the respective entities.

2.1.1. Data Term

The data term of the random field model (1) evaluates the compatibility of the proposed parameters and the observed images. In particular, it implements the assumption that corresponding image locations should be similar in appearance across the four input images. This similarity assumption is enforced by penalizing the dissimilarity between a segment in the reference view and its projection to all three remaining images. The necessary transformations are given by the combination of shape parameters and the rigid body transformation, inherited from the assigned object.

The data term depends on information from both types of hidden variables (i.e., geometry and motion). It is defined as a a pairwise potential between segments and objects

$$\varphi_i(\mathbf{s}_i, \mathbf{o}) = \sum_{k \in \mathcal{O}} [l_i = k] \cdot D_i(\mathbf{n}_i, \mathbf{o}_k) \tag{2}$$

where $[\cdot]$ denotes the Iverson bracket which ensures that φ_i is only evaluated with respect to the currently assigned object. The function $D_i(\mathbf{n}_i, \mathbf{o}_k)$ denotes a dissimilarity measure for superpixel $\mathbf{s}_i$ that depends on plane parameters $\mathbf{n}_i$ and the rigid body motion of the assigned object $\mathbf{o}_k$. To gather information from all images, the dissimilarity measure is composed of a stereo, a flow and a cross term. The three terms are computed between a reference view and the other three images, as illustrated in Figure 1a. Without loss of generality, we define the left image at t_0 as the reference view. The

complete dissimilarity measure in the data term reads as follows:

$$D_i(\mathbf{n}, \mathbf{o}) = D_i^{\text{stereo}}(\mathbf{n}, \mathbf{o}) + D_i^{\text{flow}}(\mathbf{n}, \mathbf{o}) + D_i^{\text{cross}}(\mathbf{n}, \mathbf{o}) \tag{3}$$

Each of the constituting terms is defined in (4) as the sum of matching costs C over all pixels $\mathbf{p}$ inside superpixel s_i. Matching costs are computed by transforming each pixel according to a homography induced by the associated geometry and motion. The comparison of image sites around the reference pixel and the transformed target pixel can be expressed as:

$$D_i^{\text{x}}(\mathbf{n}, \mathbf{o}) = \sum_{\mathbf{p} \in \mathcal{R}_i} C_{\text{x}}(\mathbf{p}, \underbrace{\mathbf{K}\left(\mathbf{R}_{\text{x}}(\mathbf{o}) - \mathbf{t}_{\text{x}}(\mathbf{o}) \cdot \mathbf{n}^T\right)\mathbf{K}^{-1}}_{3\times3 \text{ homography } \mathbf{H}_{\text{x}}} \mathbf{p}) \tag{4}$$

Here, x refers to one of the different matching modalities specified above: x $\in$ {stereo, flow, cross}. $\mathbf{K} \in \mathbb{R}^{3\times3}$ denotes the camera calibration matrix containing the elements of the interior orientation. For clarity of presentation, the interior orientation of the left and right camera is assumed to be equal. The transformation parameters $(\mathbf{R}_{\text{x}}(\mathbf{o}), \mathbf{t}_{\text{x}}(\mathbf{o}))$ are applied to map a 3D point in reference coordinates to a 3D point in another camera coordinate system according to the relative camera orientation and the rigid body motion of $\mathbf{o}_k$. To directly transform homogeneous image coordinates from one view to another a two-dimensional projective transformation is applied. The corresponding 3×3 homography matrix $\mathbf{H}_{\text{x}}$ is composed of two planar projections. First, the reference pixel is transformed to a three-dimensional object point in the plane of its superpixel. Next, it is mapped to the target image plane. For the stereo term the original plane parameters are used while for projections to images at t_1 the plane normal is transformed according to the object motion. Consequently, $\mathbf{R}_{\text{x}}$ and $\mathbf{t}_{\text{x}}$ depend on the matching modality x and are augmented with the parameters of the relative camera orientation where necessary.

The matching cost $C_{\text{x}}(\mathbf{p}, \mathbf{q})$ returns a dissimilarity measure 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 the proposed model, we take advantage of dense correspondences as well as sparsely matched image features. Matching costs $C_{\text{x}}(\mathbf{p}, \mathbf{q})$ are defined as a weighted sum of these two groups of observations with individual weights θ:

$$C_{\text{x}}(\mathbf{p}, \mathbf{q}) = \theta_{1,\text{x}}\, C_{\text{x}}^{\text{dense}}(\mathbf{p}, \mathbf{q}) + \theta_{2,\text{x}}\, C_{\text{x}}^{\text{sparse}}(\mathbf{p}, \mathbf{q}) \tag{5}$$

The dense matching cost $C_{\mathrm{x}}^{\mathrm{dense}}(\mathbf{p}, \mathbf{q})$ is computed via the Hamming distance (denoted by $\|\cdot\|_{\mathrm{h}}$ in the following) of the respective 5×5 Census descriptors:

$$C_{\mathrm{x}}^{\mathrm{dense}}(\mathbf{p}, \mathbf{q}) = \begin{cases} \rho_{C_{\max}}\left(\|I^C(\mathbf{p}) - I_{\mathrm{x}}^C(\mathbf{q})\|_{\mathrm{h}}\right) & \text{if } \mathbf{q} \in \Omega \\ C_{\mathrm{out}} & \text{otherwise} \end{cases} \tag{6}$$

I^C, I_{x}^C denote the Census transformed versions of the reference view and the respective target image and we introduce $\mathbf{q}$ as a shorthand for the transformed pixel position $\mathbf{H}_{\mathrm{x}}\mathbf{p}$. This patch-based similarity measure was introduced by Zabih & Woodfill (1994) and found to work well in the context of optical flow estimation (Vogel et al., 2013a). It efficiently builds a descriptor of a small image region by concatenating the binary results of intensity value comparisons. Consequently, it is robust against additive changes in illumination. An outlier value of C_{out} is assigned to points leaving the image domain Ω.

While the employed Census descriptor accounts for small deviations from the similarity assumption, it cannot cope with systematic influences like occlusions or perspective distortion. To limit the effect of such grossly wrong observations, e.g. next to depth discontinuities, a robust penalty function is $\rho_\tau(x)$ is applied to compute the matching cost. It truncates the distance x at threshold τ: $\rho_\tau(x) = \min(x, \tau)$. An overview of all truncation parameters is provided in Table 1.

In addition to the dense matching term, a second type of observation is exploited. It evaluates the consistency of displacements induced by the estimated parameters and those computed by specialized large-displacement matching approaches:

$$C_{\mathrm{x}}^{\mathrm{sparse}}(\mathbf{p}, \mathbf{q}) = \begin{cases} \rho_{\tau_1}\left(\|\pi_{\mathrm{x}}(\mathbf{p}) - \mathbf{q}\|_2\right) & \text{if } \mathbf{p} \in \Pi_{\mathrm{x}} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Here, $\pi_{\mathrm{x}}(\mathbf{p})$ denotes the transformation of pixel $\mathbf{p}$ according to the respective sparse feature correspondence, $\mathbf{q}$ is the result of transforming the reference pixel according to the estimated parameters as before. Π_{x} is the set of pixels in the reference image for which correspondences have been established. Again, x refers to the matching modality and the truncation threshold τ_1 limits the influence of outliers in the observations. Details about the employed matching approaches will be given in Section 4.

2.1.2. Smoothness Term

The task of the smoothness term in (1) is to encourage smooth transitions between adjacent superpixels. The smoothness term decomposes into three parts, weighted by parameters θ:

$$\psi_{ij}(\mathbf{s}_i, \mathbf{s}_j) = \theta_3\, \psi_{ij}^{\text{depth}}(\mathbf{n}_i, \mathbf{n}_j) + \theta_4\, \psi_{ij}^{\text{orientation}}(\mathbf{n}_i, \mathbf{n}_j) + \theta_5\, \psi_{ij}^{\text{motion}}(\mathbf{s}_i, \mathbf{s}_j) \tag{8}$$

First, regularization of depth is achieved by penalizing different disparity values d at shared boundary pixels $\mathcal{B}_{ij}$:

$$\psi_{ij}^{\text{depth}}(\mathbf{n}_i, \mathbf{n}_j) = \sum_{\mathbf{p} \in \mathcal{B}_{ij}} \rho_{\tau_2} \left(\| d(\mathbf{n}_i, \mathbf{p}) - d(\mathbf{n}_j, \mathbf{p}) \|_1 \right) \tag{9}$$

Here, the function $d(\mathbf{n}, \mathbf{p})$ returns the disparity at pixel $\mathbf{p}$ induced by the plane normal $\mathbf{n}$ of the respective segment. This well established constraint is derived from the observation that surfaces of distinct objects typically exhibit only gradual changes of geometry. To allow for depth discontinuities, as typically encountered at object boundaries, the robust penalty function ρ_{τ_2} from the previous section is applied.

Second, the orientation of neighboring planes is encouraged to be similar. This is a necessary extension of the preceding pairwise term to fully formalize the aforementioned smoothness assumption. While the requirement of consistent disparity at boundary pixels attaches neighboring segments, it does not penalize implausible folds in the reconstructed surface. Thus, our second term evaluates the similarity of plane normals $\mathbf{n}$:

$$\psi_{ij}^{\text{orientation}}(\mathbf{n}_i, \mathbf{n}_j) = \rho_{\tau_3} \left(1 - \frac{|\mathbf{n}_i^T \mathbf{n}_j|}{(\|\mathbf{n}_i\| \|\mathbf{n}_j\|)} \right) \tag{10}$$

Again, a threshold is applied to allow for sudden changes of surface orientation where needed.

Finally, we encourage coherence of the assigned object indices by an orientation-sensitive Potts model:

$$\psi_{ij}^{\text{motion}}(\mathbf{s}_i, \mathbf{s}_j) = w(\mathbf{n}_i, \mathbf{n}_j) \cdot [l_i \neq l_j] \tag{11}$$

The intuition behind this term is to penalize fragmented objects by adding a penality wherever neighboring superpixels are assigned to different objects. The weight is

defined as

$$w(\mathbf{n}_i, \mathbf{n}_j) = \frac{|\mathbf{n}_i^T \mathbf{n}_j|}{(\|\mathbf{n}_i\|\|\mathbf{n}_j\|)} \cdot \exp\left(-\frac{\alpha}{|\mathcal{B}_{ij}|} \sum_{\mathbf{p} \in \mathcal{B}_{ij}} (d(\mathbf{n}_i, \mathbf{p}) - d(\mathbf{n}_j, \mathbf{p}))^2\right) \tag{12}$$

and prefers motion boundaries that coincide with folds in the reconstructed surface. Here, α is the shape parameter of the penalty function, which is normalized by the number of shared boundary pixels $|\mathcal{B}_{ij}|$. Furthermore, the penalty increases if the surface orientation of the compared superpixels is similar.

2.2. Joint Estimation of Vehicles and Scene Flow

The scene flow model introduced in the previous section is based on a decomposition of the observed scene into rigidly moving objects. So far, rigidity is the major assumption made with respect to the objects. In this section, the flexibility of the developed scene flow model is demonstrated. In particular, by adding a semantic interpretation to the motion-based segmentation, additional shape knowledge can be incorporated into the model. In the following, we focus on the perception of traffic scenes as encountered in the context of autonomous driving and introduce a parametric shape model of cars, which we infer jointly with the scene flow and the segmentation. On the one hand, dedicated object models allow for an efficient parametrized reconstruction of highly relevant parts of the scene. The accurate estimation of object pose and shape establishes the basis for further analyses. On the other hand, high-level object knowledge can support regularization of the ill-posed scene flow problem.

2.2.1. 3D Object Model

More specifically, we leverage the Active Shape Model from Zia et al. (2013) to encode the geometry of the objects $\mathbf{o}$ introduced in Section 2.1. A training set of 38 manually annotated CAD models of passenger cars forms the basis for this geometric representation. Principal component analysis is applied to a set of manually annotated key points to retrieve the directions of the most dominant deformations between the samples in the training set. Based on the resulting Active Shape Model, novel object instances can be generated within the range of deformations in the training set, see Figure 2 for an illustration.

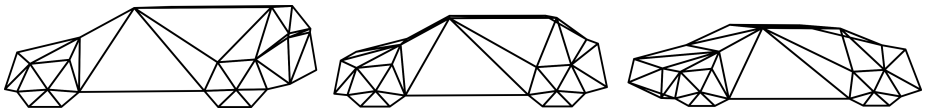


Figure 2: **Instantiations of our 3D Active Shape Model.** The mean shape is shown in the center with $\boldsymbol{\gamma} = (0, 0)$. The left and right panels contain instances illustrating the range of possible deformations with shape parameters $\boldsymbol{\gamma}_{\text{left}} = (-1.0, -0.8)$ and $\boldsymbol{\gamma}_{\text{right}} = (1.0, 0.8)$.

To incorporate shape and pose parameters into the scene flow approach, the basic object model $\mathbf{o}$, introduced in Section 2.1, is extended by two additional vectors. The shape parameters $\boldsymbol{\gamma}$ control the influence of the individual deformations of the ASM. Further, the vector $\boldsymbol{\xi}$ comprises the pose parameters of the extended object model. The two-dimensional position on the ground plane and a heading angle provide a compact representation of the position and orientation of the model relative to the reference camera. The extended representation of objects $(\boldsymbol{\gamma}_k, \boldsymbol{\xi}_k, \mathbf{R}_k, \mathbf{t}_k)$ comprises a total of 11 parameters for foreground objects (6 for rigid body motion, 3 for pose and 2 for shape).

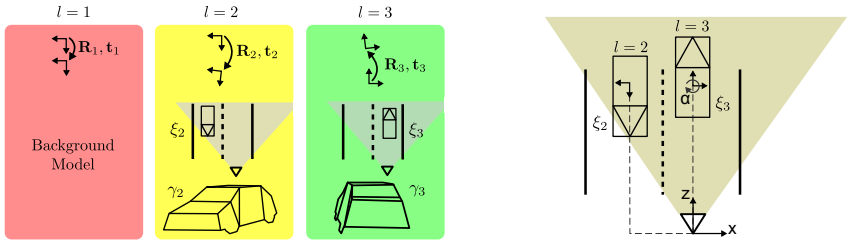
We jointly infer these variables with the scene flow and segmentation as shown in Figure 3. In panel (a), the red box represents the static background and the yellow box corresponds to the object in Figure 1b. In this example, the green box represents an additional spurious motion hypothesis, which is not associated with any of the superpixels. This underlines the fact that the proposed model is capable of performing implicit model selection (i.e., it is able to also determine the *number* of rigidly moving components) while only an upper bound on the number of expected rigid body motions is required. Panel (b) illustrates the pose parameters ξ that define the position and orientation of the objects.

2.2.2. Extension of the Scene Flow Model

To constrain the additional parameters of the revised object model, an additional shape and pose consistency term is incorporated into our random field formulation:

$$E(\mathbf{s}, \mathbf{o}) = \sum_{i \in \mathcal{S}} \Big(\underbrace{\varphi_i(\mathbf{s}_i, \mathbf{o})}_{\text{data}} + \underbrace{\kappa_i(\mathbf{s}_i, \mathbf{o})}_{\text{shape\&pose}} \Big) + \sum_{(i,j) \in \mathcal{N}} \underbrace{\psi_{ij}(\mathbf{s}_i, \mathbf{s}_j)}_{\text{smoothness}} \tag{13}$$

Here, $\mathbf{o}$ is the extended object representation introduced in the previous section, $\mathbf{s}$ represents the same planar superpixels as before and $\mathcal{N}$ denotes the set of adjacent super-



(a) Transformation and Shape $\{(\mathbf{R}_k, \mathbf{t}_k, \boldsymbol{\gamma}_k)\}_{k=1}^{|O|}$ (b) Object Pose ($\boldsymbol{\xi}_k$)

Figure 3: **Object Representation.** (a) Each object, including the static background, is represented by its rigid body motion $(\mathbf{R}_k, \mathbf{t}_k)$. Additionally, object pose ξ_k and shape γ_k can be inferred for each foreground object. (b) The object pose ξ_k is defined as the 2D position in the ground plane (x, z) and a heading angle α of the object.

pixels.

The novel shape and pose consistency term $\kappa(\cdot)$ encourages the set of estimated 3D object models to agree with the planes of the associated superpixels. In accordance with the employed parametrization of scene flow, the evaluation of a consistency measure is conducted in disparity space. Like the original data term in equation (2), the additional term κ is composed of computationally tractable pairwise potentials between superpixels and the assigned objects:

$$\kappa_i(\mathbf{s}_i, \mathbf{o}) = \theta_6 \sum_{k \in O} [l_i = k] \cdot S_i(\mathbf{n}_i, \mathbf{o}_k) + [l_i \neq k \wedge k > 1] \cdot O_{ik}(\mathbf{o}_k) \tag{14}$$

Here, κ_i defines the cost function for a distinct image segment $\mathbf{s}_i$. It combines the shape consistency term S and an occlusion penalty O, which will be defined in the following.

The shape consistency term $S_i(\mathbf{n}_i, \mathbf{o}_k)$ enforces consistency between the shape of object $\mathbf{o}_k$ and the assigned planes described by $\mathbf{n}_i$. In analogy with the data term, shape consistency is evaluated with respect to the object associated with the superpixel via l_i. The penalty function S_i considers two cases

$$S_i(\mathbf{n}, \mathbf{o}) = \begin{cases} C^{\text{bg}} & \text{if } \mathbf{o} \text{ is background} \\ \sum_{\mathbf{p} \in \mathcal{R}_i} C^{\text{obj}}(\mathbf{n}, \mathbf{o}, \mathbf{p}) & \text{otherwise} \end{cases} \tag{15}$$

C^{bg} denotes a constant penalty for superpixels associated with the background object and is imposed to avoid a bias towards purely static scenes. $C^{\text{obj}}(\mathbf{n}, \mathbf{o}, \mathbf{p})$ compares two

disparity maps and evaluates the shape consistency between superpixel and foreground object parameters at all pixels covered by the superpixel, i.e. $\mathbf{p} \in \mathcal{R}_i$, as follows:

$$C^{\text{obj}}(\mathbf{n}, \mathbf{o}, \mathbf{p}) = \begin{cases} \rho_{\theta_7}(|d(\mathbf{n}, \mathbf{p}) - d(\mathbf{o}, \mathbf{p})|) & \text{if } \mathbf{p} \in \pi(\mathbf{o}) \\ \theta_7 & \text{otherwise} \end{cases} \tag{16}$$

Here, $\pi(\mathbf{o})$ denotes the set of pixels covered by the projection of an object $\mathbf{o}$ onto the image plane and $d(\mathbf{o}, \mathbf{p})$ returns the disparity induced by the projection of object $\mathbf{o}$ at pixel $\mathbf{p}$. The pixel-wise penalty C^{obj} is computed as the truncated absolute difference (i.e., $\rho_{\theta_7}(x) = \min(x, \theta_7)$) between the virtual disparity from the projection of $\mathbf{o}$ and the disparity induced by the plane $\mathbf{n}$. Differences are computed for all pixels which coincide with the projection of $\mathbf{o}$. Pixels remaining uncovered by the projected model are penalized with θ_7, a multiple of C^{bg}. This encourages the projected model to approximately align with superpixel boundaries. Note that in contrast to the data term D_i, which encourages consistency between estimated 3D plane parameters and image observations, this term evaluates the consistency between the deformed 3D shape model and the reconstructed superpixels.

The second part of (14) is the occlusion penalty, which is formally defined as

$$O_{ik}(\mathbf{o}_k) = \theta_{\text{occ}} \cdot \sum_{\mathbf{p} \in \mathcal{R}_i} [\mathbf{p} \in \pi(\mathbf{o}_k)] \tag{17}$$

It penalizes overlap between parts of a foreground model and superpixels that are assigned to a different object via the arguments of the leading Iverson bracket in (14). We found this term crucial to prevent object models from exceeding the true object boundaries.

2.3. Preprocessing & Initialization

In the previous sections, we specified our scene flow model which contains a mixture of discrete and continuous variables: While the plane normals $\mathbf{n}_i$ of image segments and the motion parameters $(\mathbf{R}_k, \mathbf{t}_k)$ of objects live in continuous domains, the label l_i corresponds to a discrete object index. As optimizing a joint discrete-continuous energy function is hard, we iteratively discretize the continuous variables and solve a sequence of discrete energy minimization problems to find an approximate solution.

In this section, we describe the pre-processing of the data and the initialization of the parameters.

2.3.1. Image Segments and Correspondences

We first segment the reference view into superpixels and calculate disparity maps of both stereo pairs using SPS-Stereo (Yamaguchi et al., 2014), which is a state-of-the-art approach to these problems. According to the paper, images from the KITTI dataset are decomposed into approximately one thousand segments, which is shown to yield best performance concerning stereo matching and flow estimation.

In addition, we compute semi-dense associations between the reference view and the subsequent frame of the left camera via Discrete Flow (Menze et al., 2015a). The method leverages discrete optimization techniques to accurately and efficiently estimate large-displacement optical flow. While for the stand-alone optical flow method the resulting matches are interpolated using EpicFlow (Revaud et al., 2015), we directly use the sparse, discrete matches as observations in equation (7).

Both types of observations are combined to establish correspondences connecting the reference view to the right frame at t_1, i.e., optical flow vectors in the left image are combined with disparities in the left view at t_1 to yield an initial prediction of the complete displacement vector for the cross term.

2.3.2. Scene Flow Model

Next, we extract object hypotheses based on the sparse observations described above. Motion segmentation is applied to detect individually moving objects in the initial scene flow field computed via SPS-Stereo and Discrete Flow as described above.

More specifically, the sparse approximation of the three-dimensional displacement field is examined to reveal consistent motion patterns. First, we recover the dominant motion in the scene. It typically describes the relative motion between background and camera. We apply a standard approach to robust rigid motion estimation from two stereo image pairs as described in Geiger et al. (2011). The resulting transformation parameters $(\mathbf{R}, \mathbf{t})$ define the background motion hypothesis and can be leveraged for motion-based segmentation of the remaining objects.

Before extracting objects we reduce the number of outliers in the correspondence set by excluding image regions for which the scene flow leaves the image domain in any of the other views. More specifically, all pixels with valid disparities in the reference view are triangulated and projected into the target view based on the background motion. Static points falling outside the image domain are removed from the set of sparse correspondences.

Next, we detect all foreground object hypotheses as consistent clusters of scene flow correspondences that disagree with the background motion. More specifically, a threshold of 5 pixels is applied to the endpoint error of motion vectors induced by the background motion and the sparse scene flow correspondences. The retained displacement vectors are clustered according to their rigid body motion. From the set of outliers, we randomly sample 50 initial correspondences throughout the image. A three-dimensional rigid motion transformation is robustly fit to all correspondences within a sphere of radius 2.5 meters (approximating an average car's volumetric extent) around the initial matches using the 3-point RANSAC algorithm. Subsequently, we retrieve the top $|O| - 1$ hypotheses while applying non-maximum suppression to avoid multiple overlapping proposals corresponding to the same object. Note that false positive object hypotheses are typically not critical as they will not be associated with any of the image segments during inference.

2.3.3. 3D Object Model

To initialize the parameters of the extended scene flow model, we adapt the motion-based segmentation described in the previous section to the more complex parametrization of this model. First, each object hypothesis is initialized by the mean shape of the ASM (i.e., $\boldsymbol{\gamma} = 0$). We use the mean 3D coordinates of object hypotheses, reduced to the ground plane, as approximate values for the position. To complete the initial object pose parameters $\boldsymbol{\xi}$ we compute the heading angle of the objects from their moving directions. For approximately symmetric objects, like cars, it is important to carefully choose initial values as to avoid failure cases, like 180° turns, due to this ambiguity. Working with individually moving objects, this information can be extracted reliably from the moving direction. As before, this initialization procedure will lead to some

spurious object hypotheses. During inference, such false positives are pruned if no superpixels are associated with them.

2.4. Inference

Given the initialization of variables described in the previous section, we now aim to minimize the objective function (1). Inspired by the work of Yamaguchi et al. (2012), we apply particle max-product belief propagation (MP-PBP) (Trinh & McAllester, 2009), which iteratively resamples all continuous variables and solves a sequence of discrete optimization problems, minimizing the objective function (1) in each iteration. In this section, we first describe the details of our inference algorithm for the basic scene flow model and then explain how to incorporate the 3D model based extension.

2.4.1. Scene Flow Model

In each iteration of MP-PBP, we sample a set of hypotheses around the current solution of each parameter and select the proposal which minimizes the objective function to perform derivative-free optimization. We include the best solution so far into the proposal set to ensure that the objective function decreases at each iteration. More specifically, proposals are drawn from normal distributions around the current estimate. The variance of the distributions is reduced after each iteration to refine the discretization and encourage convergence. Additionally, the proposal set of each superpixel is augmented by a fixed number of MAP solutions from neighboring image sites, which increases diversity in the proposal set and propagates promising proposals to nearby image locations.

Note that even after discretization, optimization of the loopy CRF specified in (1) with respect to all superpixel and object parameters is an NP-hard combinatorial problem. We therefore compute an approximate solution at each MP-PBP iteration via sequential tree-reweighted message passing (TRW-S) (Kolmogorov, 2006).

2.4.2. 3D Object Model

Sampling the object shape and pose parameters required for the extended model in addition to the rigid body transformation would result in a large number of particles, significantly increasing the computational complexity of the problem. To keep

computation tractable, we perform "informed" sampling of pose and shape parameters based on the respective data term. In each iteration of the outer loop, we first draw 50 particles representing object pose and shape from normal distributions centered at the preceding MAP solution. As before, the respective standard deviations are iteratively reduced. To prune the proposals, the shape and pose consistency term (14) is evaluated for each particle with respect to the disparity map induced by the current MAP solution of the superpixels. Only the best shape/pose particle of each object is accepted. During the optimization of the remaining parameters, the shape and pose consistency term remains active, guiding the optimization of the remaining variables.

3. KITTI 2015 Scene Flow Dataset

For evaluating our scene flow approach with respect to competing methods, an appropriate benchmark dataset is required. Unfortunately, the creation of reference data for motion fields is a challenging task on its own. One reason for this is that there exists virtually no sensor capable of capturing ground truth correspondences in challenging real-world scenes, leading to a shortage of appropriate reference data, especially for the task of scene flow evaluation. As an alternative to real data, synthetic renderings of spheres (Huguet & Devernay, 2007; Valgaerts et al., 2010), other geometric primitives (Vogel et al., 2011; Cech et al., 2011; Basha et al., 2013) or simple street scenes (Wedel et al., 2008; Rabe et al., 2010) have typically been employed to measure quantitative performance.

Departing from this paradigm, the KITTI benchmark suite provided the first realistic platform to evaluate stereo and motion algorithms (Geiger et al., 2012), providing a range of challenges with a focus on automotive applications. The provided stereoscopic image sequences were captured from a car driving in regular traffic on public roads. Three-dimensional reference data has been captured by a 360° laser scanner mounted on top of the car. A similar approach has been used by Kondermann et al. (2015) who register stereo imagery with scans of an urban environment. Although the images are much more realistic compared to synthetic renderings, both datasets provide reference data for static scenes only. This prohibits the evaluation of the core properties

(a) Large Displacements

(b) Cast Shadows on Dynamic Objects

(c) Low Light Conditions

Figure 4: **Challenges of the proposed KITTI 2015 Scene Flow Evaluation.**

of any scene flow algorithm, namely to precisely estimate the dynamic nature of the scene and the individual motion of independently moving objects therein.

We therefore augmented the KITTI vision benchmark with a scene flow extension (Menze & Geiger, 2015) capturing dynamic scenes. In this article, this extension will be referred to as "KITTI 2015" as opposed to the original "KITTI 2012" stereo and flow benchmark. Our extension comprises very challenging outdoor scenes with depth and motion ground truth even for individually moving objects, making it the first re-

alistic scene flow dataset. Figure 4 provides an overview over the challenges present in our dataset, including (a) large displacements, (b) severe shadows, and (c) low light conditions. All of these adverse effects are commonplace in automotive applications and pose difficulties to current image matching algorithms.

Starting from the raw data, which was collected for the KITTI project (Geiger et al., 2013), we selected 200 training and 200 test images with large independent motions based on the annotated 3D object trajectories in the KITTI raw data. Our annotation process consists of two major steps that are explained in the following. First, the static background of the scene is recovered from laser range measurements. Second, the dynamic elements in the scene are annotated with the help of 3D CAD models.

For recovering the static background and registering it to the moving camera, the laser scans are first corrected with respect to 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 yielded 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. For each reference view, we accumulate scans over a temporal window of 7 frames in a common coordinate system. We further remove all 3D points belonging to moving objects using the 3D bounding box annotations provided on the KITTI website[2].

Unfortunately, the dynamic elements in the scene cannot be recovered from sparse and noisy 3D laser measurements alone. For this reason, we inserted detailed 3D CAD models of cars from 3D Warehouse[3]. It is important to note that, given the limited measurement accuracy of stereo techniques, our 3D CAD models are not required with millimeter-accuracy, which would be intractable considering the broad variety of vehicles in the video footage. Instead, we select the most similar model from a limited but diverse set of vehicles from 3D Warehouse. For each model, we obtain a three-dimensional point cloud by uniformly sampling approximately $3,000$ points from all

[2] `http://www.cvlibs.net/datasets/kitti/raw_data.php`

[3] `https://3dwarehouse.sketchup.com/`

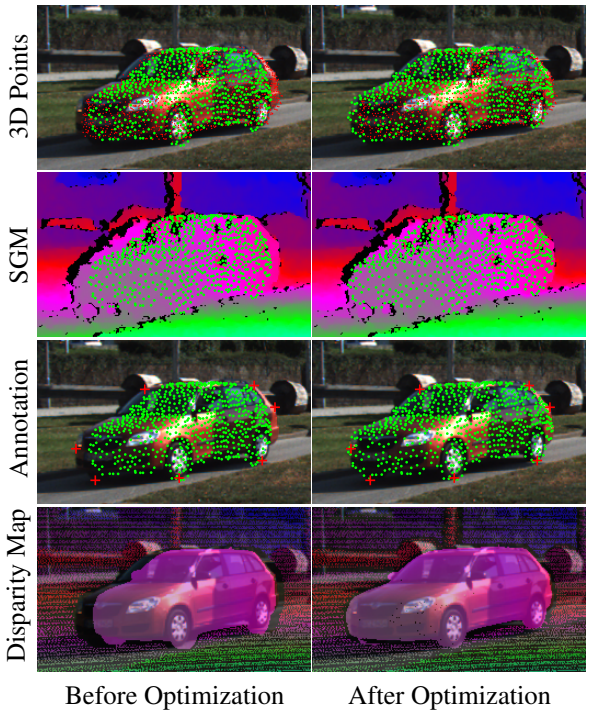


Figure 5: **Annotation process.** The topmost panel of this figure shows the subsampled CAD model (green) together with the 3D laser observations (red) used for registering the model. Additionally, we use disparity maps (second row) and 2D manual annotations (red crosses, third row) to optimize the model. The last row shows the resulting disparity map. All panels are split to show the state before (left) and after (right) minimizing equation (18).

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. Figure 5 provides an overview of the annotation process.

More specifically, for each dynamic object in the scene, we estimate a 3D similarity transformation defining the 3D 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 $\zeta \in \mathbb{R}^{15}$ (three for translation, three rotation angles, three for scaling in the reference view and six rigid motion parameters). 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 semiglobal matching. In contrast to the laser points, these observations are not subject to the rolling shutter effect. 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 the results. As a third observation, we introduce manually annotated correspondences between geometrically meaningful vertices 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.

Given these observations, we obtain the transformation parameters ζ by minimizing the following energy function

$$E(\zeta) = \sum_{t\in\{1,2\}} \left(\theta_{3D}\ E_t^{3D}(\zeta)\right)^2 + \left(\theta_{SGM}\ E_t^{SGM}(\zeta)\right)^2 + \left(\theta_{2D}\ E_t^{2D}(\zeta)\right)^2 \tag{18}$$

where t is the frame index and E_t are the energy terms corresponding to each of the observations. More specifically, E_t^{3D} denotes the truncated distance between the set of 3D laser points $\mathcal{P}$ inside the object's 3D bounding box and their geometrically nearest neighbors in the transformed CAD model:

$$E_t^{3D}(\zeta) = \sum_{\mathbf{p}\in\mathcal{P}} \rho_{\tau_{3D}}\left(\|\mathbf{X}_{CAD}(\zeta) - \mathbf{X}_{\mathbf{p}}\|_2\right) \tag{19}$$

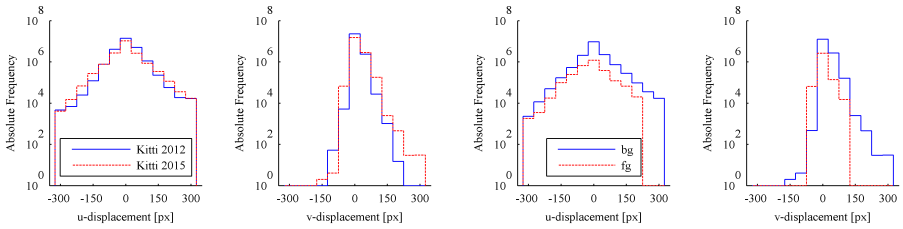
E_t^{SGM} represents the truncated distance between the disparity map induced by the transformed CAD model and the set of valid SGM measurements $\mathcal{D}$ covered by the model in image space:

$$E_t^{SGM}(\zeta) = \sum_{\mathbf{p}\in\mathcal{D}} \rho_{\tau_{SGM}}\left(|d_{SGM}(\zeta,\mathbf{p}) - d(\mathbf{p})|\right) \tag{20}$$

E_t^{2D} is the error with respect to the manually selected $2D-3D$ correspondences C in frame t. In accordance with the annotations it is computed in image space with respect to the projected CAD vertices $\mathbf{X}_{CAD}^{2D}$:

$$E_t^{2D}(\zeta) = \sum_{\mathbf{p}\in C} \|\mathbf{X}_{CAD}^{2D}(\zeta) - \mathbf{p}\|_2 \tag{21}$$

Our optimization scheme alternates between minimizing equation (18) with respect to ζ using non-linear least-squares estimation and updating all nearest neighbor associations until convergence. The weights of the terms are chosen to ensure a dominating influence of the manual input.



(a) KITTI 2012 vs. KITTI 2015 (b) KITTI 2015: Background vs. Foreground

Figure 6: **Histograms of Displacements in u- and v-axis.** Panel (a) compares the displacements contained in KITTI 2015 and KITTI 2012 benchmarks. Panel (b) shows the displacements on foreground and background objects, which are only relevant for KITTI 2015. For the sake of clarity, the y-axes are scaled logarithmically.

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 ζ. To handle occlusions within and between objects, we employ a z-buffer to decide which points are used in the reference data. Finally, non-rigidly moving objects like pedestrians or bicyclists and erroneous regions in the laser scans are masked manually. Histograms of the resulting displacements are provided in Figure 6. 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 degrade the accuracy of the ground truth. All results are therefore evaluated using a dedicated metric that takes these error characteristics into account (see Section 4).

The dataset and an online evaluation on test data with held back reference are available as part of the KITTI benchmark suite[4]. In addition to the evaluation of scene flow estimates, it allows for the individual evaluation of results for the stereo matching and optical flow sub-problems.

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

4. Evaluation

This section provides a detailed experimental evaluation of the proposed approach. We start with details on the scene flow evaluation metric and a brief overview of our training strategy. Next, we report quantitative results of our model and compare the performance of the proposed method to the state-of-the-art. Finally, we present a number of qualitative results and analyze the abilities and limitations of the approach.

4.1. Evaluation Protocol

Our novel KITTI 2015 scene flow benchmark evaluates four values per pixel which uniquely determine the 3D scene flow field: Two disparity values, one in the first and one in the second frame, and the displacement in x and y direction. To ease evaluation and mitigate the problem of occlusions, all values are represented in the reference image coordinate system, e.g., the disparity value of the second frame (after displacement) is stored at the corresponding pixel location before displacement. Our dedicated scene flow metric evaluates the combination of all three measures, i.e., only pixels with correct disparities *and* flow are considered as correct scene flow estimates. All entities are evaluated at each valid ground truth pixel in the reference view. An estimate is considered wrong if one of the disparity values or the optical flow vector exceeds a distance of 3 pixels *and* 5% of the respective ground truth value. This combination of error metrics ensures an evaluation which is faithful with respect to the uncertainties in the reference data (due to the complex annotation process, large displacements are assigned a tolerance with regard to their magnitude). Summary statistics over all 200 test images are collected by averaging errors over valid reference values of foreground and background regions and the combination of both.

4.2. Parameter Training

The scene flow model described in Section 2 contains a number of parameters, which can be trained to adapt the model to specific datasets. To enable unbiased comparison and to decrease the computational burden during training, the available training data is split into two sets of equal size. One half is used for parameter training and the other serves validation purposes. Due to the small number of hyper-parameters in the

Parameter	Value	Parameter	Value
$\theta_{1,\text{stereo}}$	1.00	θ_7, θ_{occ}	4.00
$\theta_{1,\text{flow}}$	1.00	C_{bg}	1.00
$\theta_{1,\text{cross}}$	1.00	C_{max}	0.79
$\theta_{2,\text{stereo}}$	0.02	C_{out}	0.36
$\theta_{2,\text{flow}}$	0.76	$\tau_{1,\text{stereo}}$	1.82
$\theta_{2,\text{cross}}$	0.76	$\tau_{1,\text{flow}}$	3.90
θ_3	0.38	$\tau_{1,\text{cross}}$	3.90
θ_4	14.79	τ_2	2.56
θ_5	83.13	τ_3	0.26
θ_6	0.30	α	0.20

Table 1: **Model Parameters.** This table provides the model parameter values, trained as described in Section 4.2.

model, we estimated their values on a subset of the KITTI 2015 training set using a discrete variant of block coordinate descent. This simple, iterative strategy discretizes the solution space of each variable within a plausible range. The training error is computed for each of the samples and the best-performing parameter value is chosen. To account for some of the correlations, for example between truncation thresholds and weights, some of the parameters are jointly optimized. This strategy implies quadratic growth of the number of samples in parameter space and is thus restricted to the most obvious dependencies. The parameter values provided in Table 1 are the result of 10 block coordinate descent iterations and are used throughout all of our experiments.

4.3. Quantitative Results

The quantitative analysis of the scene flow approach consists of two major parts. First, the importance of the model components is investigated in ablation studies. Second, the full scene flow model and the extension to 3D model reconstruction are compared numerically to the results of state-of-the-art scene flow methods.

#		D1			D2			Fl			SF		
		bg	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*
1	Data (SPSS+SpF)	3.8	12.4	5.2	4.8	21.4	7.4	5.5	25.1	8.5	6.7	30.4	10.4
2	Data (Census)	4.8	12.8	6.0	5.8	14.4	7.1	6.1	19.0	8.1	7.6	25.1	10.3
3	Data (All)	3.9	11.0	5.0	4.8	12.9	6.1	5.4	17.4	7.3	6.6	23.0	9.1
4	Data (All) + Smooth (Boundary)	**3.5**	**8.6**	**4.3**	**4.4**	10.7	5.4	5.1	16.1	6.8	6.2	20.3	8.4
5	Data (All) + Smooth (Normal)	3.8	10.7	4.9	4.7	12.6	6.0	5.4	17.2	7.2	6.5	22.6	9.0
6	Data (All) + Smooth (Object)	4.0	12.1	5.3	4.9	13.2	6.2	5.4	17.2	7.2	6.6	22.7	9.1
7	Data (SPSS+SpF) + Smooth (All)	4.6	15.1	6.2	5.7	16.1	7.3	6.2	18.1	8.0	7.6	23.1	10.0
8	Data (Census) + Smooth (All)	3.6	8.9	4.4	4.5	**10.0**	5.4	5.1	15.5	6.7	6.2	19.4	8.2
9	Data (All) + Smooth (All)	**3.5**	9.2	4.4	**4.4**	10.6	**5.3**	**5.0**	**15.1**	**6.5**	**6.0**	**19.2**	**8.1**

Table 2: **Influence of Scene Flow Model Components.** This table shows the error rates for disparities in the reference frame (D1) and the target frame (D2), optical flow (Fl) and scene flow (SF) averaged over all 100 validation images. For each modality, the outlier percentage is reported for the background region (bg), all foreground objects (fg) as well as all annotated pixels in the image (bg&fg). The evaluation is conducted on the validation portion of the KITTI 2015 training set.

4.3.1. Ablation Studies

First, we assess the contribution of each individual term in the energy function (1). Our evaluation is conducted on a validation portion of 100 training images from the scene flow dataset. Table 2 shows the results when evaluating all annotated image locations. The columns show errors in terms of disparity at both time steps ("D1", "D2"), optical flow ("Fl") and scene flow ("SF") using the conventions specified in Section 4.1. For each modality, the table provides results in terms of the static background ("bg"), individually moving foreground objects ("fg") as well as the combination of both ("bg&fg"). The first three rows of the table show the results of Object Scene Flow using only the data terms. The overall scene flow error is comparable when using only sparse *or* dense features, and is reduced significantly using the combination of both. Rows 4 to 6 show results for different combinations of data terms and selected smoothness terms. It can be seen that the boundary term in row 4 is the strongest pairwise cue. In combination with all data terms, it produces the lowest error rates for disparities in the first frame. The remaining pairwise terms encourage consistently moving objects and contribute to D2, Fl and SF. Again, the combination of all pairwise terms, shown in row 9, yields the overall best scene flow results. The last three rows show the two

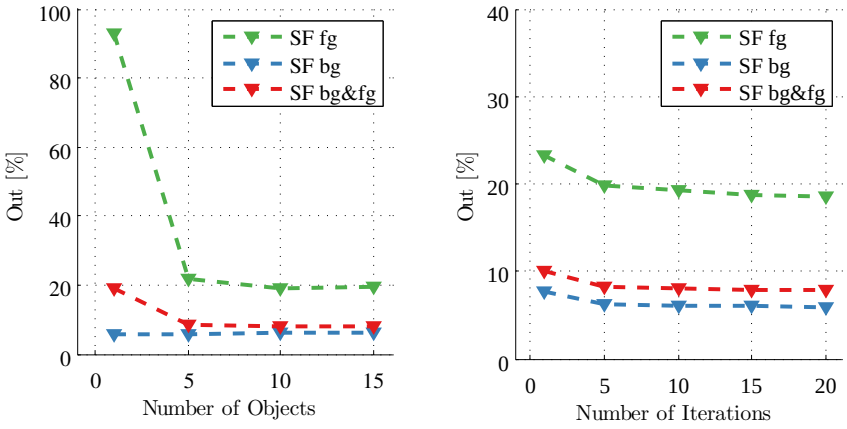


Figure 7: **Model Parameters.** This figure shows the percentage of scene flow outliers of our method with respect to the number of object proposals and iterations of the particle-based optimization. Different colors encode the results for foreground regions (green), background regions (blue) and the combined metric (red). The evaluation is conducted on the validation portion of the KITTI 2015 training set.

groups of data terms together with all smoothness terms and the full model. In combination with the pairwise terms, the dense Census features are almost on par with the full model, which yields the overall best results.

Next, we investigate the performance of the full model with respect to the size of the object set and the number of iterations, see Figure 7. In particular, we vary the number of allowed object hypotheses in our model between 1 and 15. Note that by allowing only one object we restrict the model to completely static scenes, effectively implementing a motion stereo approach. The left panel of Figure 7 affirms our assumption that the outdoor scenes we consider can be described sufficiently well by a small number of rigidly moving objects. The overall error (shown in red) drops significantly from 1 to 5 objects. It improves slightly to its minimal value at 10 objects and then starts to rise again. A moderate number of ten objects accounts for two phenomena: On the one hand, it covers complex scenes with many visible objects and distinct motions. On the other, it ensures a large enough number of object hypotheses to allow for some false detections from the motion-based segmentation and still cover the true objects. As can be expected, the background error is not affected by the number of

objects since it typically corresponds to the dominant object in the scene. The right panel of Figure 7 shows the performance of our method with respect to the number of iterations in the particle-based optimization framework. This plot shows that the error rate reduces significantly within the first 5 iterations and then saturates at 10 iterations.

Based on these results we chose the parameters for the following experiments. We use 10 shape particles per superpixel, 10 objects and 5 motion particles per object. As a tradeoff between run time and overall accuracy 10 iterations of max-product particle belief propagation (MP-PBP) are performed. All motion particles and half of the shape particles are drawn from a normal distribution centered at the MAP solution of the last iteration or the initialization, respectively. The remaining shape particles are proposed using the plane parameters from spatially neighboring superpixels. These are randomly sampled conditioned on the distance of superixel centers. Both strategies complement each other and we found their combination important for efficiently exploring the search space.

4.3.2. Comparison to the state-of-the-art

Table 3 compares the error rates of the proposed Object Scene Flow (*OSF 2018*) to several baselines on all annotated image locations of the KITTI 2015 test data. The table contains all methods which were submitted and published by the end of 2016. To ensure a fair comparison based on all annotated pixels, the results of sparse and semi-dense methods are interpolated using a standard routine provided by the KITTI development kit.

Besides the classic variational approach (VSF) of Huguet & Devernay (2007) and the recent continuous method (CSF) of Lv et al. (2016), the table provides results for the prediction-correction approach (PCOF) of Derome et al. (2016) and the sparse scene flow method of Cech et al. (2011) (GCSF). For the initial release of the benchmark, further baselines were constructed by combining two state-of-the-art optical flow algorithms with disparity estimates in both frames obtained using semiglobal matching. In particular, we combine SGM with large displacement optical flow (LDOF) (Brox & Malik, 2011) and with a classical hierarchical variational approach with nonlocal regularization (C+NL) (Sun et al., 2014). As a representative for RGB-D based

	D1			D2			Fl			SF			Run
	bg	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	time [s]
PRSM (multi-frame)	3.02	10.52	4.27	5.13	15.11	6.79	5.33	13.40	6.68	6.61	20.79	8.97	300
OSF 2018	**4.11**	**11.12**	**5.28**	**5.01**	**17.28**	**7.06**	**5.38**	**17.61**	**7.41**	**6.68**	**24.59**	**9.66**	390
OSF 2015	4.54	12.03	5.79	5.45	19.41	7.77	5.62	18.92	7.83	7.01	26.34	10.23	3000
OSF-M 2018	4.47	11.62	5.66	5.69	18.30	7.79	6.03	19.66	8.29	7.53	26.25	10.65	500
CSF	4.57	13.04	5.98	7.92	20.76	10.06	10.40	25.78	12.96	12.21	33.21	15.71	80
PRSF	4.74	13.74	6.24	11.14	20.47	12.69	11.73	24.33	13.83	13.49	31.22	16.44	150
SGM+SF	5.15	15.29	6.84	14.10	23.13	15.60	20.91	25.50	21.67	23.09	34.46	24.98	2700
PCOF-LDOF	6.31	19.24	8.46	19.09	30.54	20.99	14.34	38.32	18.33	25.26	49.39	29.27	50
PCOF + ACTF (GPU)	6.31	19.24	8.46	19.15	36.27	22.00	14.89	60.15	22.43	25.77	67.75	32.76	0.08
SGM+C+NL	5.15	15.29	6.84	28.77	25.65	28.25	34.24	42.46	35.61	38.21	50.95	40.33	270
SGM+LDOF	5.15	15.29	6.84	29.58	23.48	28.56	40.81	31.92	39.33	43.99	42.09	43.67	86
DWBSF	19.61	22.69	20.12	35.72	28.15	34.46	40.74	31.16	39.14	46.42	40.76	45.48	420
GCSF	11.64	27.11	14.21	32.94	35.77	33.41	47.38	41.50	46.40	52.92	56.68	53.54	3
VSF	27.31	21.72	26.38	59.51	44.93	57.08	50.06	45.40	49.28	67.69	62.93	66.90	7500

Table 3: **Results on the proposed KITTI 2015 Test Set.** This table shows error rates for disparities in the reference frame (D1) and the target frame (D2), optical flow (Fl) and scene flow (SF) averaged over all 200 test images. For each modality, the outlier percentage is reported for the background region (bg), all foreground objects (fg) as well as all annotated pixels in the image (bg&fg). The table contains all methods which were submitted and published by the end of 2016.

algorithms, the results of Sphere Flow (SGM+SF) by Hornacek et al. (2014) are provided. To emulate the required depth component 3D object points were reconstructed from all valid pixels of the SGM disparity maps. The results of the piece-wise rigid scene flow (PRSF) approach by Vogel et al. (2013b) were computed with the original parameter setting, which was trained on KITTI 2012.

Table 3 contains a duplicate entry of OSF in the third row, entitled OSF 2015, that provides the results of the first version published in Menze & Geiger (2015). In contrast to the variant described in this paper disparity observations were originally computed using SGM instead of SPS-Stereo and the sparse optical flow matches from Geiger et al. (2011) were used instead of Discrete Flow. The improved observations used in this work allow for a reduced number of shape particles and iterations during inference. A decrease of the outlier percentage in all evaluated categories is accompanied by a significant reduction of the required run time. Overall, the proposed method yields top performance amongst all two-frame scene flow methods to date on the challeng-

ing KITTI 2015 evaluation. However, while leading to plausible qualititative results, the model-based extension (OSF-M 2018) does not lead to increased quantitative performance. We analyze the reasons for this in the experimental evaluation. Further, the advantages of processing more than two frames is demonstrated by PRSM (Vogel et al., 2014). It obtains the best results amongst all published methods to date, but outperforms our two-frame scene flow method only slightly. We expect similar gains in the results of our method when extending it to more than two frames.

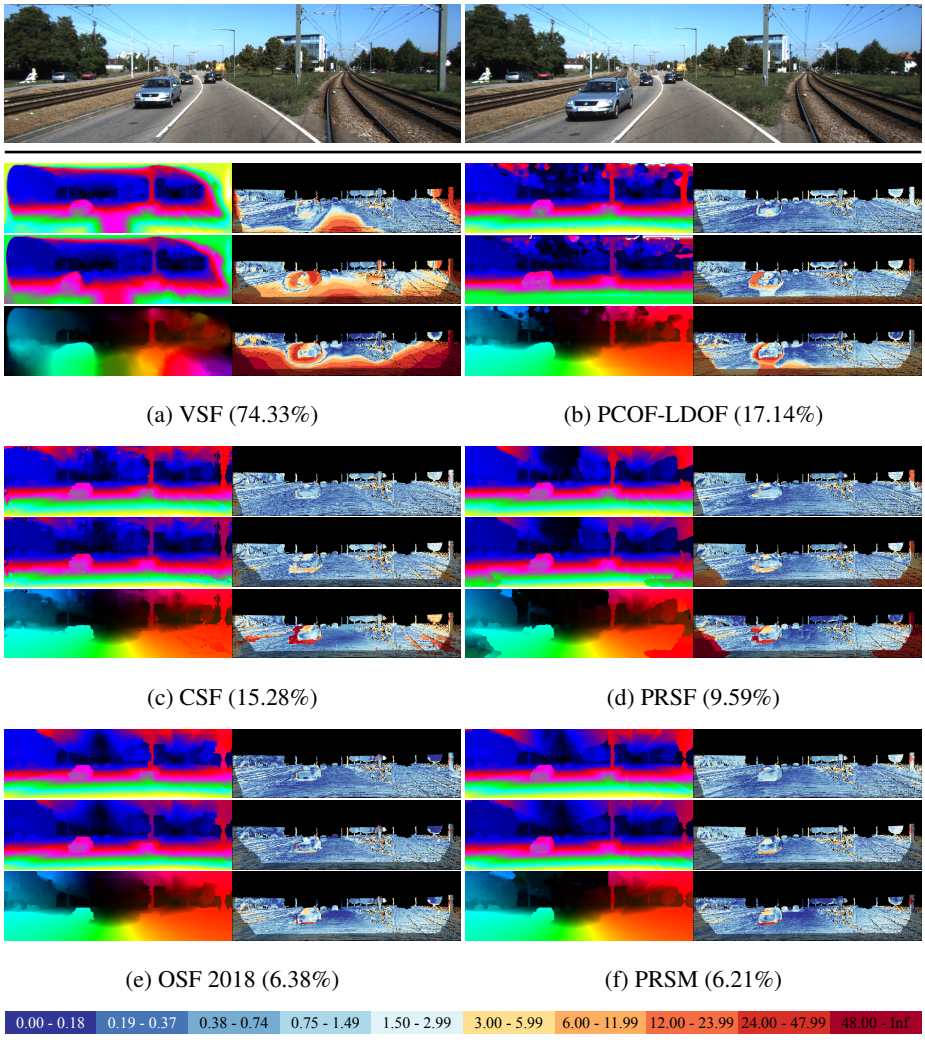
4.4. Qualitative Results

In the following, we provide qualitative results to asses the strengths and weaknesses of the proposed method. In Figure 8, we present the results of our approach compared to competing submissions for one example sequence. Figure 9 shows additional results of our method, including the model-based extension.

Figure 8 shows the results as published on the KITTI 2015 website at the time of submission (February 2017). In particular, we compare the results of the proposed OSF model with respect to a variational baseline (VSF) and other top-performing methods (PRSF, PRSM) on the benchmark. The top row shows the input images of the left camera. Due to the strong relative motion, the car on the left-most lane undergoes significant perspective distortion and induces heavy occlusions. The results are ordered with respect to descending overall scene flow error on the provided example. The subfigures below the input images are split into visualizations of disparity in both frames (first and second row) and the estimated optical flow map (third row), shown in the left part, as well as color-coded error maps, shown on the right. The legend for the error maps is provided at the bottom of the figure. Inliers according to the combined 3px / 5% scene flow metric are depicted in blue shades while outliers are depicted in red shades.

The optical flow field estimated by VSF in panel (a) is only correct near the center of the image where the observed displacements are small. In contrast, large displacements are successfully recovered by the prediction-correction approach PCOF-LDOF, shown in panel (b). It provides significantly smaller errors in the flow map but around the vehicle it suffers from erroneous results in the disparity map at t_2. The contin-

uous method CSF in panel (c) optimizes a slanted-plane model using a variational framework. While CSF improves upon PCOF-LDOF, some of the very small image segments result in gross errors due to the difficulties of CSF in estimating the optical flow correctly.



(a) VSF (74.33%) (b) PCOF-LDOF (17.14%)

(c) CSF (15.28%) (d) PRSF (9.59%)

(e) OSF 2018 (6.38%) (f) PRSM (6.21%)

Figure 8: **Comparison to Related Work.** The top row provides the input imagery from the left camera. Each sub-figure shows from top-to-bottom: The estimated disparity maps at t_1 and t_2 and the estimated optical flow between both views (left panel) as well as the respective error images using the logarithmic color scheme depicted in the legend (right panel). Inliers are depicted in blue shades while outliers are depicted in red shades.

The occluded region left of the car on the opposite lane is also difficult for PRSF. As a consequence, the scene flow error is slightly bigger than that of the proposed OSF, which handles the car successfully up to small bleeding artifacts due to imperfect segmentation. The benefits of taking into account more than two input frames are evidenced by the results of PRSM in panel (f).

Figure 9 provides additional qualitative results of our method and the model-based extension. Here, each subfigure shows (from top-to-bottom) the disparity (left) and optical flow (right) ground truth, the first disparity map and the optical flow map estimated by our method and the respective error maps in the last row. The lower part of the figure additionally shows the predicted models before (first row) and after (second row) inference overlayed as wireframe models onto the disparity and optical flow estimates.

The first row of this figure contains two examples on which the proposed method works well. As evidenced by the error images, it is able to recover the correct disparity and optical flow in a variety of challenging situations. The effect of sub-optimal segmentation is evidenced around the outline of foreground objects and on the traffic lights in both depicted scenes, causing bleeding artifacts around the cars. However, most of the annotated disparities and displacements on the narrow poles are recovered correctly. In contrast to the related approach of Vogel et al. (2013b), we do not refine the initial segmentation during inference. This helps to limit the computational burden. As a consequence, errors in the segmentation directly transfer to artifacts at object boundaries. To counter this effect, superpixels are chosen to be sufficiently small to faithfully capture scene geometry and motion. Including the segmentation into the optimization will be an interesting avenue for future research.

We observed that even objects that are not perfectly rigid are detected and assigned plausible estimates. As an example consider the bicyclist in the top-left panel of Figure 9, which moves at the same speed as the observing vehicle. Therefore, the optical flow map shows brighter colors indicating smaller displacements compared to the surrounding background. However, as neither the laser scanner nor the rigid CAD models provide appropriate annotations for articulated objects, we excluded the respective image regions manually from the quantitative evaluation in the KITTI 2015 benchmark.

In the right panel of the first row, the proposed method struggles with background areas close to the image boundaries. The reason for this is that the leftmost region leaves the image domain of the three remaining frames. In this case, the smoothness terms are not able to extrapolate the disparity estimate correctly. The challenges at the bottom right of the image are more subtle. Here, a small stretch of grass stands out from the otherwise flat ground. The rightmost portion of it is supported by enough image evidence to induce a depth jump, while the rest is smoothed by the slanted-plane model. Overall, however, the scene flow error in these two examples lies below the test set average reported in Table 3.

The second row of Figure 9 provides results with higher error than on average. In the left panel, some gross errors are evident on the vegetation area next to the road. Small twigs and leaves contradicting the slanted plane assumption are responsible for these errors. They are caused by the chosen model but, in many applications, such very fine details can be considered irrelevant. On the more important persistent parts of the scene, like the traffic sign and the visible tree trunks, OSF yields mostly correct disparities and displacements. The disparity map on the left shows a faithful segmentation of the delivery truck. While the reconstruction is largely successful, the optical flow result is less accurate around the roof of the vehicle.

The example on the right of the second row can be considered a failure case. Difficult lighting conditions, reflecting surfaces and a quickly moving object on the opposite lane render this example especially challenging. Striking errors in the disparity map occur on the leftmost vehicle and the bright area next to it. In addition, OSF fails to recover the motion of the approaching vehicles.

The lower part of Figure 9 illustrates resulting disparity and optical flow maps together with wire-frame renderings of the object models. The panels show results for six representative scenes. The top row of each sub-figure depicts the layout after initialization as described in Section 2. In most cases, the shapes do not match the observed cars and there are some significant translational offsets. In addition, there are many spurious objects initialized due to wrong object hypotheses. The center row of each panel shows our reconstruction results after optimizing the energy function (13). Objects which are not assigned to any of the superpixels are considered absent and thus

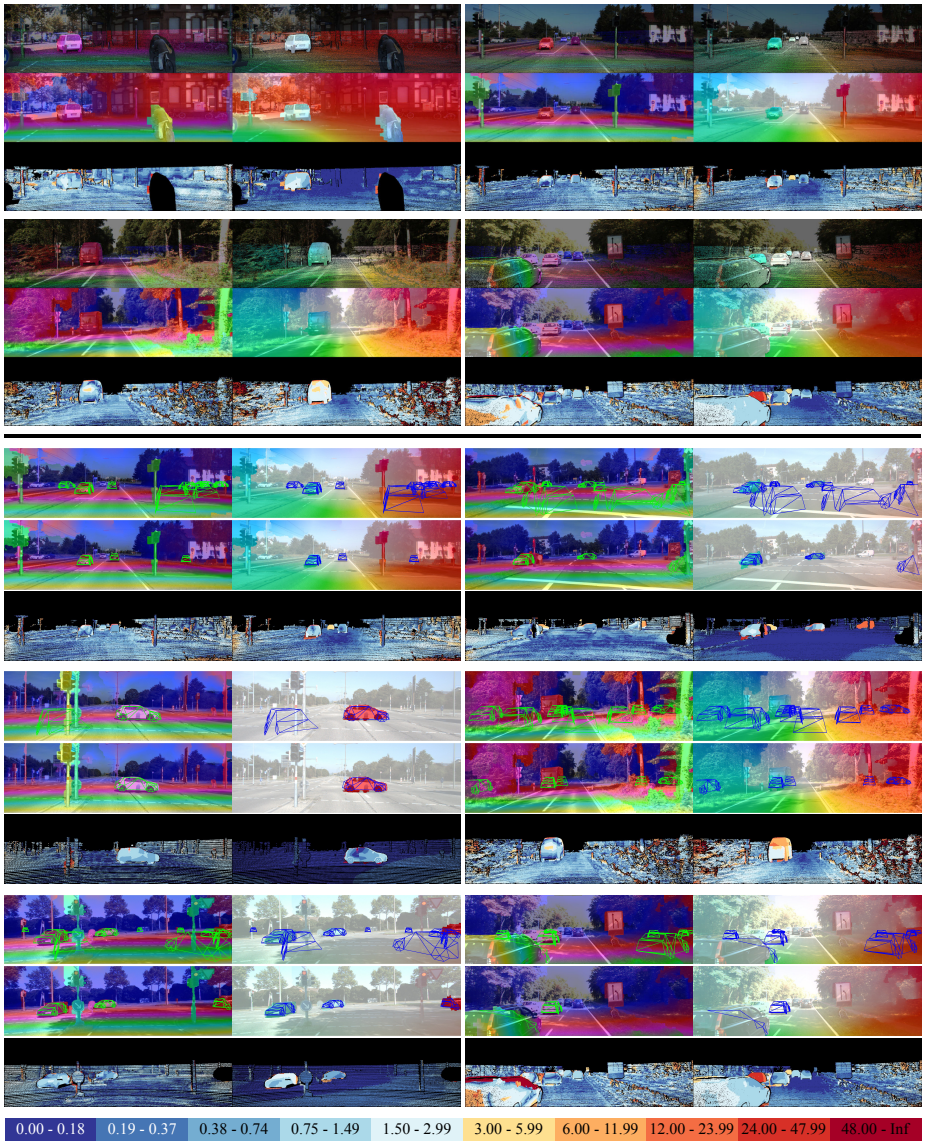


Figure 9: **Qualitative Results of our Scene Flow Approach.** In the upper part, each sub-figure shows from top-to-bottom: The disparity and optical flow ground truth in the reference view, the disparity map (D1) and optical flow map (Fl) estimated by our scene flow algorithm, and the respective error images using the logarithmic color scheme depicted in the legend. The lower part additionally shows the results of the model-based extension before and after inference.

not drawn. The last row of each panel provides color-coded error maps as before.

For all examples shown in the first row, the model position is successfully aligned with the observed object and the shape of the model is faithfully adapted to the depicted cars. Further, spurious hypotheses are initialized next to the road and on some road markings. They are successfully removed, demonstrating the intrinsic model selection capability of the approach.

In the second row, we compare a very good result to a representative failure case of our method. The left panel shows one successfully reconstructed car in the foreground. The initialization also contains one spurious object around the traffic light, which is removed during inference. On the right panel, the depicted van exceeds the shape space of the Active Shape Model. Therefore, two models are fitted to the consistently moving region corresponding to the van. In the flow error map on the right, it becomes obvious that the upper part of the vehicle is not assigned to the reconstructed objects. This is expected, as the respective image segments are not covered by the object models. Under the proposed setup, failure cases like this lead to the increased average errors in Table 3. Besides, complex scene geometry, as encountered next to the road in this example, can randomly generate groups of consistently deviating displacements. Such motion cues may cause false object hypotheses. Sometimes, these objects remain even after optimization, if the model geometry can be adapted to agree with the observations. False object hypotheses are also initialized in the left panel of the last row of Figure 9. In this case, however, the corresponding image regions are correctly assigned to the background during inference.

Containing six individually moving cars, the scene depicted in the last panel on the right is one of the most complex in the dataset. Here, only two of the six vehicles present in the scene are correctly initialized as object hypotheses. Two more vehicles are moving in the same direction as the observing car. As they are located close to the center of the image and move slowly, they are missed by the motion-based segmentation. Two vehicles on the opposite lane are missed as well. One of them appears relatively small in the input image and is likely to be missed due to larger, erroneous object hypotheses. The other is depicted in a saturated area that does not provide strong image evidence. While two of the initialized object hypotheses remain after optimization

only the central car is recovered correctly. The model for the leftmost car is initialized with a significant angular error which cannot be recovered during inference. Adding a recognition component to our approach for detecting the shape and pose of objects in the scene could help in the aforementioned situation and has already demonstrated promising results on other task, e.g., stereo estimation (Güney & Geiger, 2015). We consider the combination of scene flow and recognition as a promising direction for future work.

5. Conclusions

In this paper, we presented a novel model for joint rigid motion segmentation, 3D scene flow estimation and 3D model fitting. In addition, we developed the first realistic benchmark for 3D scene flow evaluation in challenging dynamic outdoor scenes and compared our method to the state-of-the-art techniques on this dataset. Our core technical contribution is the efficient parametrization of the problem via superpixels and individually moving objects. It maintains the necessary flexibility of the model while imposing valuable constraints by introducing long-range spatial dependencies. We experimentally demonstrated that our assumptions hold for the processed outdoor image sequences and lead to improved results on the challenging KITTI 2015 dataset. Additional performance gains can be expected when extending our approach to more than 2 frames. While the computational complexity of the resulting algorithm allows for tractable inference, it still significantly exceeds the requirements of real-time applications. We plan to address this issue in the future. Besides, we believe that a combination of the presented model with object recognition or semantic segmentation to exploit the synergistic effects between these modalities will be promising for future research.

References

Bao, S., Chandraker, M., Lin, Y., & Savarese, S. (2013). Dense object reconstruction with semantic priors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

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

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

Braun, C., Kolbe, T. H., Lang, F., Schickler, W., Steinhage, V., Cremers, A. B., Förstner, W., & Plümer, L. (1995). Models for photogrammetric building reconstruction. In *Computers & Graphics*.

Brox, T., Bruhn, A., Papenberg, N., & Weickert, J. (2004). High accuracy optical flow estimation based on a theory for warping. In *Proc. of the European Conf. on Computer Vision (ECCV)*.

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

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

Cootes, T. F., Taylor, C. J., Cooper, D. H., & Graham, J. (1995). Active shape models-their training and application. *Computer Vision and Image Understanding (CVIU)*, *61*, 38–59.

Dame, A., Prisacariu, V., Ren, C., & Reid, I. (2013). Dense reconstruction using 3D object shape priors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Debevec, P. E., Taylor, C. J., & Malik, J. (1996). Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *ACM Trans. on Graphics (SIGGRAPH)*.

Derome, M., Plyer, A., Sanfourche, M., & Le Besnerais, G. (2016). A prediction-correction approach for real-time optical flow computation using stereo. In *Proc. of the German Conference on Pattern Recognition (GCPR)*.

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

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research (IJRR)*, *32*, 1231–1237.

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

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

Güney, F., & Geiger, A. (2015). Displets: Resolving stereo ambiguities using object knowledge. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

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

Hirschmüller, H. (2008). Stereo processing by semiglobal matching and mutual information. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, *30*, 328–341.

Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence (AI)*, *17*, 185–203.

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

Huguet, F., & Devernay, F. (2007). A variational method for scene flow estimation from stereo sequences. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*.

Isack, H., & Boykov, Y. (2012). Energy-based geometric multi-model fitting. *International Journal of Computer Vision (IJCV)*, *97*, 123–147.

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

Kondermann, D., Nair, R., Meister, S., Mischler, W., Güssefeld, B., Honauer, K., Hofmann, S., Brenner, C., & Jähne, B. (2015). Stereo ground truth with error bars. In *Proc. of the Asian Conf. on Computer Vision (ACCV)*.

Leibe, B., Cornelis, N., Cornelis, K., & Van Gool, L. (2006). Integrating recognition and reconstruction for cognitive traffic scene analysis from a moving vehicle. In *Proc. of the DAGM Symposium on Pattern Recognition (DAGM)*.

Lv, Z., Beall, C., Alcantarilla, P., Li, F., Kira, Z., & Dellaert, F. (2016). A continuous optimization approach for efficient and accurate scene flow. In *Proc. of the European Conf. on Computer Vision (ECCV)*.

Menze, M., & Geiger, A. (2015). Object scene flow for autonomous vehicles. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Menze, M., Heipke, C., & Geiger, A. (2015a). Discrete optimization for optical flow. In *Proc. of the German Conference on Pattern Recognition (GCPR)*.

Menze, M., Heipke, C., & Geiger, A. (2015b). Joint 3d estimation of vehicles and scene flow. In *Proc. of the ISPRS Workshop on Image Sequence Analysis (ISA)*.

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

Prisacariu, V., Segal, A., & Reid, I. (2013). Simultaneous monocular 2d segmentation, 3d pose recovery and 3d reconstruction. In *Proc. of the Asian Conf. on Computer Vision (ACCV)*.

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

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

Revaud, J., Weinzaepfel, P., Harchaoui, Z., & Schmid, C. (2015). EpicFlow: Edge-preserving interpolation of correspondences for optical flow. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

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

Sun, D., Roth, S., & Black, M. J. (2014). A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision (IJCV)*, *106*, 115–137.

Thomas, A., Ferrari, V., Leibe, B., Tuytelaars, T., & Van Gool, L. (2007). Depth-from-recognition: Inferring meta-data by cognitive feedback. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*.

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

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

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

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

Vogel, C., Roth, S., & Schindler, K. (2013a). An evaluation of data costs for optical flow. In *Proc. of the German Conference on Pattern Recognition (GCPR)*.

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

Vogel, C., Schindler, K., & Roth, S. (2011). 3D scene flow estimation with a rigid motion prior. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*.

Vogel, C., Schindler, K., & Roth, S. (2013b). Piecewise rigid scene flow. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*.

Vogel, C., Schindler, K., & Roth, S. (2015). 3d scene flow estimation with a piecewise rigid scene model. *International Journal of Computer Vision (IJCV)*, *115*, 1–28.

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

Yamaguchi, K., Hazan, T., McAllester, D., & Urtasun, R. (2012). Continuous markov random fields for robust stereo estimation. In *Proc. of the European Conf. on Computer Vision (ECCV)*.

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

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

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

Zhang, H., Geiger, A., & Urtasun, R. (2013). Understanding high-level semantics by modeling traffic patterns. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*.

Zhou, C., Güney, F., Wang, Y., & Geiger, A. (2015). Exploiting object similarity in 3d reconstruction. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*.

Zia, M., Stark, M., Schiele, B., & Schindler, K. (2013). Detailed 3D representations for object recognition and modeling. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, *35*, 2608–2623.

Zia, M., Stark, M., & Schindler, K. (2015). Towards scene understanding with detailed 3d object representations. *International Journal of Computer Vision (IJCV)*, *112*, 188–203.