VOLDOR: Visual Odometry from Log-logistic Dense Optical flow Residuals

Zhixiang Min Yiding Yang Enrique Dunn
Stevens Institute of Technology
{zmin1,yyang99,edunn}@stevens.edu

Abstract

We propose a dense indirect visual odometry method taking as input externally estimated optical flow fields instead of hand-crafted feature correspondences. We define our problem as a probabilistic model and develop a generalized-EM formulation for the joint inference of camera motion, pixel depth, and motion-track confidence. Contrary to traditional methods assuming Gaussian-distributed observation errors, we supervise our inference framework under an (empirically validated) adaptive log-logistic distribution model. Moreover, the log-logistic residual model generalizes well to different state-of-the-art optical flow methods, making our approach modular and agnostic to the choice of optical flow estimators. Our method achieved top-ranking results on both TUM RGB-D and KITTI odometry benchmarks. Our open-sourced implementation [1] is inherently GPU-friendly with only linear computational and storage growth.

1. Introduction

Visual odometry (VO) [60, 21, 22] addresses the recovery of camera poses from an input video sequence, which supports applications such as augmented reality, robotics and autonomous driving. Traditional indirect VO [47, 64, 57] methods rely on the geometric analysis of sparse keypoint correspondences to determine multi-view relationships among the input video frames. Moreover, by virtue of relying on local feature detection and correspondence pre-processing modules, indirect methods pose the VO problem as a reprojection-error minimization task. Conversely, direct methods [58, 16, 45], strive to jointly determine a (semi-)dense registration (warping) across images, as well as the parameters of a camera motion model. By virtue of evaluating a dense correspondence field, direct methods strive to minimize photometric error among registered images. While both of these contrasting approaches have been successful in practice, there are important limitations to be addressed. An open problem within indirect methods, is how to characterize feature localization error within the context of VO [41, 40, 39, 93], where motion blur, depth occlusions and viewpoint variations may corrupt such estimates. Nevertheless, least squares methods are commonly used under the assumption of zero-mean Gaussian distributed observation errors. On the other hand, the efficacy of direct methods relies on the strict adherence to the small-motion and appearance-constancy assumptions (or on the development of registration models robust to such variations), which speaks to the difficulty of adequately modeling data variability in this context and, in turn, reduces the scope of their applicability.

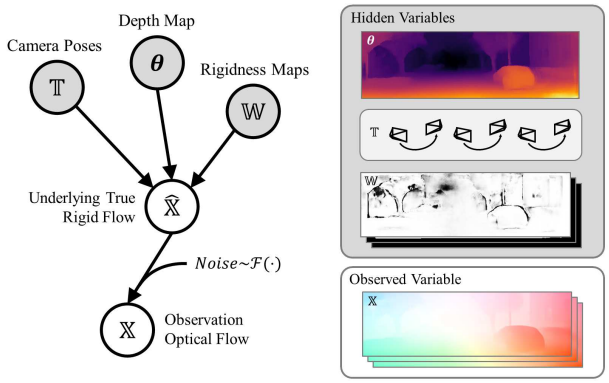


Figure 1: **VOLDOR probabilistic graphical model**. Optical flow field sequences are modeled as observed variables subject to Fisk-distributed measurement errors. Camera poses, depth map and rigidness maps are modeled as hidden variables.

Recent developments on optical flow estimation [77, 35] using supervised learning have yielded state-of-the-art performance. However, such performance benefits have not yet permeated to pose-estimation tasks, where standard multi-view geometry methods still provide the "gold standard". This work develops a dense indirect framework for monocular VO that takes as input externally computed optical flow from supervised-learning estimators. We have empirically observed that optical flow residuals tend to conform to a log-logistic (i.e. Fisk) distribution model parametrized by the optical flow magnitude. We leverage this insight to pro-

[1] https://github.com/htkseason/VOLDOR

pose a probabilistic framework that fuses a dense optical flow sequence and jointly estimates camera motion, pixel depth, and motion-track confidence through a generalized-EM formulation. Our approach is dense in the sense that each pixel corresponds to an instance of our estimated random variables; it is indirect in the sense that we treat individual pixels as viewing rays within minimal feature-based multi-view geometry models (i.e. P3P for camera pose, 3D triangulation for pixel depth) and implicitly optimize for reprojection error. Starting from a deterministic bootstrap of a camera pose and pixel depths attained from optical flow inputs, we iteratively alternate the inference of depth, pose and track confidence over a batch of consecutive images.

The advantages of our framework include: 1) We present a modular framework that is agnostic to the optical flow estimator engine, which allows us to take full advantage of recent deep-learning optical flow methods. Moreover, by replacing sparse hand-crafted features inputs by learned dense optical flows, we gain surface information for poorly textured (i.e. feature-less) regions. 2) By leveraging our empirically validated log-logistic residual model, we attain highly accurate probabilistic estimates for our scene depth and camera motion, which do not rely on Gaussian error assumptions. Experiments on the KITTI [25] and TUM RGB-D [76] benchmark have yielded top-ranking performance both on the visual odometry and depth estimation tasks. Our highly parallelizable approach also allows real-time application on commodity GPU-based architectures.

2. Related Works

Indirect methods. Indirect methods [92, 32, 59, 20, 13, 29] rely on geometric analysis of sparse keypoint correspondences among input video frames, and pose VO problem as a reprojection-error minimization task. VISO [47] employs Kalman filter with RANSAC-based outlier rejection to robustly estimate frame-to-frame motion. PTAM [64] splits tracking and mapping to different threads, and applies expensive bundle adjustment (BA) at the back-end to achieve better accuracy while retaining real-time application. ORB-SLAM [56, 57] further introduces a versatile SLAM system with a more powerful back-end with global relocalization and loop closing, allowing large environment applications.

Direct methods. Direct methods [61, 71, 98, 46, 45, 44, 70, 86] maintain a (semi-)dense model, and estimates the camera motion by finding a warping that minimizes the photometric error w.r.t. video frames. DTAM [58] introduces a GPU-based real-time dense modelling and tracking approach for small workspaces. LSD-SLAM [16] switched to semi-dense model that allows large scale CPU real-time application. DSO [15] builds sparse models and combines a probabilistic model that jointly optimizes for all parameters as well as further integrates a full photometric calibration to achieve current state-of-the-art accuracy.

Deep learning VO. Recently, deep-learning has shown thriving progress on the visual odometry problem. Boosting VO through geometric priors from learning-based depth predictions has been presented in [89, 52, 79]. Integration of deep-representations into components such as feature-points, depth maps and optimizers has been presented in [78, 5, 11, 12]. Deep-learning framework that jointly estimate depth, optical flow and camera motion has been presented in [102, 91, 94, 80, 81, 54, 103, 36, 84, 66, 72, 6, 9, 97]. Further adding recurrent neural networks for learning temporal information is presented in [82, 83, 88, 49]. However, deep learning methods are usually less explainable and have difficulty when transferring to unseen dataset or cameras with different calibration. Moreover, the precision of such methods still underperforms the state of the art.

Deep learning optical flow. Contrary to learning-based monocular depth approaches [23, 28, 27, 96, 31], which develop and impose strong semantic priors, learning for optical flow estimation may be informed by photometric error and achieve better generalization. Recent deep learning works on optical flow estimation [85, 33, 90, 65, 38, 51, 34, 101, 35, 14, 77] have shown satisfying accuracy, robustness and generalization, outperforming traditional methods, especially under challenging conditions such as texture-less regions, motion blur and large occlusions. FlowNet [14] introduced an encoder-decoder convolutional neural network for optical flow. FlowNet2 [35] improved its performance by stacking multiple basic FlowNets. Most recently, PWC-Net [77] integrates spatial pyramids, warping and cost volumes into deep optical flow estimation, improving performance and generalization to the current state-of-the-art.

3. Problem Formulation

Rationale. Optical flow can been seen as a combination of rigid flow, which is relevant to the camera motion and scene structure, along with an unconstrained flow describing general object motion [87]. Our VO method inputs a batch of externally computed optical flow fields and infers the underlying temporally-consistent aforementioned scene structure (depth map), camera motions, as well as pixel-wise probabilities for the "rigidness" of each optical flow estimate. Further, we posit the system framework under the supervision of an empirically validated adaptive log-logistic residual model over the end-point-error (EPE) between the estimated rigid flow and the input (observed) flow.

Geometric Notation. We input a sequence of externally computed (observed) dense optical flow fields $\mathbb{X} = \{\boldsymbol{X_t} \mid t = 1, \cdots, t_N\}$, where $\boldsymbol{X_t}$ is the optical flow map from image $\boldsymbol{I_{t-1}}$ to $\boldsymbol{I_t}$, while $\boldsymbol{X_t^j} = (u_t^j, v_t^j)^T$ denotes the optical flow vector of pixel j at time t. We aim to infer the camera poses $\mathbb{T} = \{\boldsymbol{T_t} \mid t = 1, \cdots, t_N\}$, where $\boldsymbol{T_t} \in SE(3)$ represents the relative motion from time $t-1$ to t.
To define a likelihood model relating our observations $\mathbb{X}$

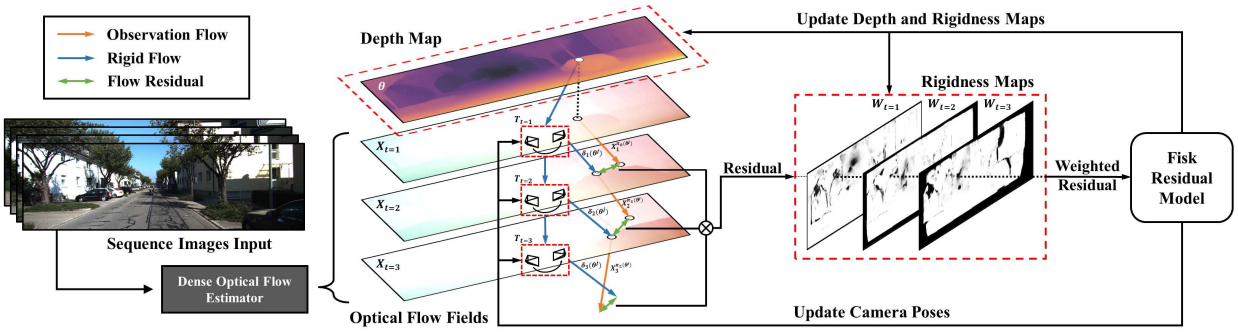


Figure 2: **Iterative estimation workflow**. We input externally computed optical flow estimates of a video sequence. Scene depth, camera poses and a rigidness maps are alternatively estimated by enforcing congruence between predicted rigid flow and input flow observations. Estimation is posed as a probabilistic inference task governed by a Fisk-distributed residual model.

to $\mathbb{T}$, we introduce two additional (latent) variable types: 1) a depth field $\boldsymbol{\theta}$ defined over $\boldsymbol{I_0}$; where we denote θ^j as the depth value at pixel j, and 2) a rigidness probability map $\boldsymbol{W_t}$ associated to $\boldsymbol{\theta}$ at time t; while $\mathbb{W} = \{\boldsymbol{W_t} \mid t = 1, \cdots, t_N\}$ denotes the set of rigidness maps, and W_t^j denotes the rigidness probability of pixel j at time t.

Having depth map $\boldsymbol{\theta}$ and rigidness maps $\mathbb{W}$, we can obtain a rigid flow $\boldsymbol{\xi_t}(\theta^j)$ by applying the rigid transformation $\mathbb{T}$ to the point cloud associated with $\boldsymbol{\theta}$, conditioned on $\mathbb{W}$. Assuming $\boldsymbol{T_0} = \boldsymbol{I}$, we let $\boldsymbol{\pi_t}(\theta^j)$ denote the pixel coordinate of projecting the 3D point associated with θ^j into the camera image plane at time t using given camera poses $\mathbb{T}$, by

$$\boldsymbol{\pi_t}(\theta^j) = \boldsymbol{K}\left(\prod_{i=0}^{t} \boldsymbol{T_i}\right)\theta^j \boldsymbol{K}^{-1}[x_j \; y_j \; 1]^T \tag{1}$$

where $\boldsymbol{K}$ is the camera intrinsic and x_j, y_j are the image coordinates of pixel j. Hence, rigid flow can be defined as $\boldsymbol{\xi_t}(\theta^j) = \boldsymbol{\pi_t}(\theta^j) - \boldsymbol{\pi_{t-1}}(\theta^j)$.

Mixture Likelihood Model We model the residuals between observation flow and rigid flow with respect to the continuous rigidness probability W_t^j.

$$P(\boldsymbol{X_t}^{\boldsymbol{\pi_{t-1}}(\theta^j)} \mid \theta^j, \boldsymbol{T_t}, W_t^j ; \boldsymbol{T_1} \cdots \boldsymbol{T_{t-1}}) = \begin{cases} \rho(\boldsymbol{\xi_t}(\theta^j) \mid\mid \boldsymbol{X_t}^{\boldsymbol{\pi_{t-1}}(\theta^j)}) & \text{if } W_t^j = 1 \\ \mu(\boldsymbol{X_t}^{\boldsymbol{\pi_{t-1}}(\theta^j)}) & \text{if } W_t^j = 0 \end{cases} \tag{2}$$

where the probability density function $\rho(\cdot \mid\mid \cdot)$ represents the probability for having the rigid flow $\boldsymbol{\xi_t}(\theta^j)$ under the observation flow of $\boldsymbol{X_t}^{\boldsymbol{\pi_{t-1}}(\theta^j)}$, and $\mu(\cdot)$ is a uniform distribution whose density varies with $\boldsymbol{X_t}^{\boldsymbol{\pi_{t-1}}(\theta^j)}$. We will define these two functions in §4. Henceforth, when modeling the probability of $\boldsymbol{X_t}$, we only write down $\boldsymbol{T_t}$ in the conditional probability, although the projection also depends on preceding camera poses $\boldsymbol{T_1}, \cdots, \boldsymbol{T_{t-1}}$, that we assume fixed and for which $\boldsymbol{X_t}$ inherently does not contain any information. Moreover, jointly modeling for all previous camera poses along with $\boldsymbol{X_t}$ would bias them as well as increase the computational complexity. In the following paragraph, we will denote Eq. (2) simply as $P(\boldsymbol{X_t}^{\boldsymbol{\pi_{t-1}}(\theta^j)} \mid \theta^j, \boldsymbol{T_t}, W_t^j)$. At this point, our visual odometry problem can be modeled as a maximum likelihood estimation problem

$$\begin{aligned} &\underset{\boldsymbol{\theta},\mathbb{T},\mathbb{W}}{argmax} P(\mathbb{X} \mid \boldsymbol{\theta}, \mathbb{T}, \mathbb{W}) \\ &= \underset{\boldsymbol{\theta},\mathbb{T},\mathbb{W}}{argmax} \prod_t \prod_j P(\boldsymbol{X_t}^{\boldsymbol{\pi_{t-1}}(\theta^j)} \mid \theta^j, \boldsymbol{T_t}, W_t^j) \end{aligned} \tag{3}$$

Furthermore, we promote spatial consistency among both our dense hidden variables θ and $\mathbb{W}$, through different mechanisms described in §5.1.

4. Fisk Residual Model

The choice of an observation residual model plays a critical role for accurate statistical inference from a reduced number of observations [30, 37], where in our case, a residual is defined as the end-point error in pixels between two optical flow vectors.

In practice, hierarchical optical flow methods (i.e. relying on recursive scale-space analysis) [77, 50] tend to amplify estimation errors in proportion to the magnitude of the pixel flow vector. In light of this, we explore an adaptive residual model determining the residuals distribution w.r.t. the magnitude of optical flow observations. In Figure 3, we empirically analyzed the residual distribution of multiple leading optical flow methods [77, 35, 50, 67] w.r.t. groundtruth. We fit the empirical distribution to five different analytic models, and found the Fisk distribution to yield the most congruent shape over all flow magnitudes (See Figure 3 overlay). To quantify the goodness of fit, in

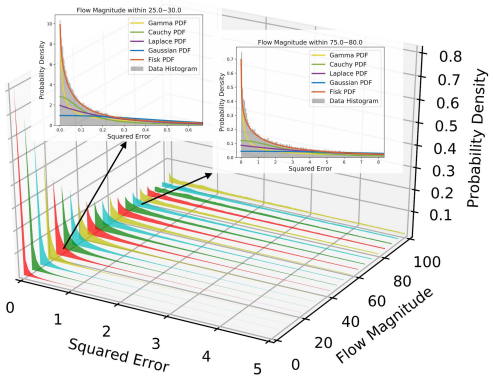
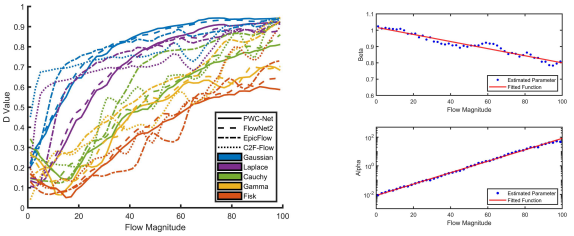


Figure 3: **Empirical residual distribution.** Optical flow EPE residual over flow magnitude for PWC-Net outputs on the entire groundtruth data for the KITTI [25] and Sintel [7] datasets.



(a) K-S Test (b) Parameterization of α, β

Figure 4: **Fitness quantification and model parameterization.** (a) Results of KS-test [3] over four optical flow methods with five distributions (lower D values indicate better fit). (b) α, β are estimated from KITTI empirical data using, respectively, log-linear and linear regression, as described in Eq. (6), (7).

Figure 4 (a), we posit K-S test results [3], which quantify the supremum distance (D value) of the CDF between our empirical distribution and a reference analytic distribution.

Hence, given $\boldsymbol{v_{ob}} = \boldsymbol{X_t^{\pi_{t-1}}}(\theta^j)$, we model the probability of $\boldsymbol{v_{rig}} = \boldsymbol{\xi_t}(\theta^j)$ matching the underlying groundtruth, as

$$\rho(\boldsymbol{v_{rig}} \| \boldsymbol{v_{ob}}) = \mathcal{F}(\|\boldsymbol{v_{rig}} - \boldsymbol{v_{ob}}\|_2^2; A(\boldsymbol{v_{ob}}), B(\boldsymbol{v_{ob}})), \quad (4)$$

where the functional form for the PDF of the Fisk distribution $\mathcal{F}$ is given by

$$\mathcal{F}(x; \alpha, \beta) = \frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{(1+(x/\alpha)^{\beta})^2} \quad (5)$$

From Figure 4 (b), we further determine the parameters of Fisk distribution. Since a clear linear correspondence is shown in the figure, instead of using a look-up table, we applied the fitted function to find the parameters, as

$$A(\boldsymbol{v_{ob}}) = a_1 \boldsymbol{e}^{a_2 \|\boldsymbol{v_{ob}}\|_2} \quad (6)$$

$$B(\boldsymbol{v_{ob}}) = b_1 \|\boldsymbol{v_{ob}}\|_2 + b_2 \quad (7)$$

where a_1, a_2 and b_1, b_2 are learned parameters depending on the optical flow estimation method.

Next, we model the outlier likelihood function $\mu(\cdot)$. The general approach [53, 99] is to assign outliers with uniform distribution to improve the robustness. In our work, for utilizing the prior given by the observation flow, we further let the density of uniform distribution be a function $\mu(\cdot)$ over the observation flow vector.

$$\mu(\boldsymbol{v_{ob}}) = \mathcal{F}(\lambda^2 \|\boldsymbol{v_{ob}}\|_2^2; A(\boldsymbol{v_{ob}}), B(\boldsymbol{v_{ob}})) \quad (8)$$

where λ is a hyper-parameter adjusting the density, which is also the strictness for choosing inliers. The numerical interpretation of λ is the optical flow percentage EPE for indistinguishable outlier ($W_t^j = 0.5$). Hence, flows with different magnitudes can be compared under a fair metrics to be selected as an inlier.

5. Inference

In this section, we introduce our iterative inference framework which alternately optimizes depth map, camera poses and rigidness maps.

5.1. Depth and Rigidness Update

Generalized Expectation-Maximization (GEM). We infer depth $\boldsymbol{\theta}$ and its rigidness $\mathbb{W}$ over time while assuming fixed known camera pose $\mathbb{T}$. We approximate the true posterior $P(\mathbb{X} \mid \boldsymbol{\theta}, \mathbb{W}; \mathbb{T})$ through a GEM framework [73]. In this section, we will denote Eq. (2) as $P(\mathbb{X} \mid \boldsymbol{\theta}, \mathbb{W})$, where the fixed $\mathbb{T}$ is omitted. We approximate the intractable real posterior $P(\boldsymbol{\theta}, \mathbb{W} \mid \mathbb{X})$ with a restricted family of distributions $q(\boldsymbol{\theta}, \mathbb{W})$, where $q(\boldsymbol{\theta}, \mathbb{W}) = \prod_j q(\theta^j) \prod_t q(\boldsymbol{W_t})$. For tractability, $q(\theta^j)$ is further constrained to the family of Kronecker delta functions $q(\theta^j) = \delta(\theta^j = \theta^{j*})$, where θ^{j*} is a parameter to be estimated. Moreover, $q(\boldsymbol{W_t})$ inherits the smoothness defined on the rigidness map $\boldsymbol{W_t}$ in Eq. (11), which is proved in [73] as minimizing KL divergence between the variational distribution and the true posterior. In the M step, we seek to estimate an optimal value for θ^j given an estimated PDF on W_t^j. Next, we describe our choice for the estimators used for this task.

Maximimum Likelihood Estimator (MLE). The standard definition of the MLE for our problem is given by

$$\theta^j_{\text{MLE}} = \underset{\theta^{j*}}{argmax} \sum_t q(W_t^j) \log P(\boldsymbol{X_t^{\pi_{t-1}}}(\theta^j) \mid \theta^j = \theta^{j*}, W_t^j) \quad (9)$$

where $q(W_t^j)$ is the estimated distribution density given by E step. However, we empirically found the MLE criteria to be overly sensitive to inaccurate initialization. More specifically, we bootstrap our depth map using only the first optical flow and use its depth values to sequentially bootstrap subsequent camera poses (more details in §5.2 and §5.3). Hence, for noisy/inaccurate initialization, using MLE for estimate refinement will impose high selectivity pressure on

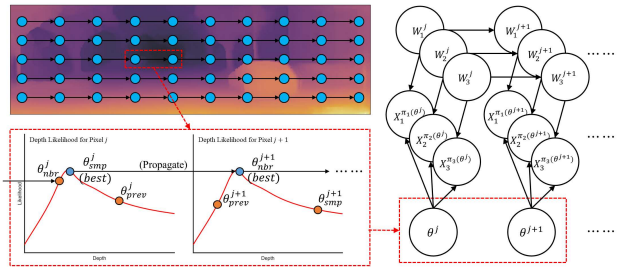


Figure 5: **Model for depth inference**. The image 2D field is broken into alternatively directed 1D chains, while depth values are propagated through each chain. Hidden Markov chain smoothing is imposed on the rigidness maps.

the rigidness probabilities $\mathbb{W}$, favoring a reduced number of higher-accuracy initialization. Given the sequential nature of our image-batch analysis, this tends to effectively reduce the set of useful down-stream observations used to estimate subsequent cameras.

Maximum Inlier Estimation (MIE). To reduce the bias caused by initialization and sequential updating, we relax the MLE criteria to the following MIE criteria,

$$\theta^j_{\text{MIE}} = \underset{\theta^{j*}}{argmax} \sum_t q(W^j_t = 1) \log \frac{P(\boldsymbol{X}_{\boldsymbol{t}}^{\boldsymbol{\pi_{t-1}}(\theta^j)} | \theta^j = \theta^{j*}, W^j_t = 1)}{\sum_{W^j_t} P(\boldsymbol{X}_{\boldsymbol{t}}^{\boldsymbol{\pi_{t-1}}(\theta^j)} | \theta^j = \theta^{j*}, W^j_t)} \tag{10}$$

which finds a depth maximizing for the rigidness (inlier selection) map $\mathbb{W}$. We provide experimental details regarding the MIE criteria in §6.3.

We optimize θ^j_{MIE} through a sampling-propagation scheme as shown in Fig. 5. A randomly sampled depth θ^j_{smp} is compared with the previous depth value θ^j_{prev} together with a value propagated from the previous neighbor θ^{j-1}_{nbr}. Then, θ^j_{MIE} will be updated to the value of the best estimation among these three options. The updated θ^j_{MIE} will further be propagated to the neighbor pixel $j+1$.

Updating the Rigidness Maps. We adopt a scheme where the image is split into rows and columns, reducing a 2D image to several 1D hidden Markov chains, and a pairwise smoothness term is posed on the rigidness map

$$P(W^j_t \mid W^{j-1}_t) = \begin{pmatrix} \gamma & 1-\gamma \\ 1-\gamma & \gamma \end{pmatrix} \tag{11}$$

where γ is a transition probability encouraging similar neighboring rigidness. In the E step, we update rigidness maps $\mathbb{W}$ according to $\boldsymbol{\theta}$. As the smoothness defined in Eq. (11), the forward-backward algorithm is used for inferring $\mathbb{W}$ in the hidden Markov chain.

$$q(W^j_t) = \frac{1}{A} m_f(W^j_t) m_b(W^j_t) \tag{12}$$

where A is a normalization factor while $m_f(W^j_t)$ and $m_b(W^j_t)$ are the forward and backward message of W^j_t computed recursively as

$$m_f(W^j_t) = P^{t,j}_{ems} \sum\nolimits_{W^{j-1}_t} m_f(W^{j-1}_t) P(W^j_t \mid W^{j-1}_t) \tag{13}$$

$$m_b(W^j_t) = \sum\nolimits_{W^{j+1}_t} m_b(W^{j+1}_t) P^{t,j+1}_{ems} P(W^{j+1}_t \mid W^j_t) \tag{14}$$

where P^j_{ems} is the emission probability refer to Eq. (2), $P^{t,j}_{ems} = P(\boldsymbol{X}_{\boldsymbol{t}}^{\boldsymbol{\pi_{t-1}}(\theta^j)} \mid \theta^j, W^j_t ; \mathbb{T})$.

5.2. Pose Update

We update camera poses while assuming fixed known depth $\boldsymbol{\theta}$ and rigidness maps $\mathbb{W}$. We use the optical flow chains in $\mathbb{X}$ to determine the 2D projection of any given 3D point extracted from our depth map. Since we aim to estimate relative camera motion we express scene depth relative to the camera pose at time $t-1$ and use the attained 3D-2D correspondences to define a dense PnP instance [100]. We solve this instance by estimating the mode on an approximated posterior distribution given by Monte-Carlo sampling of the pose space through (minimal) P3P instances. The robustness to outliers correspondences and the independence from initialization provided by our method is crucial to bootstrapping our visual odometry system §5.3, where the camera pose need to be estimated from scratch with an uninformative rigidness map (all one).

The problem can be written down as maximum a posterior (MAP) by

$$\underset{\mathbb{T}}{argmax} P(\mathbb{T} \mid \mathbb{X} ; \boldsymbol{\theta}, \mathbb{W}) \tag{15}$$

Finding the optimum camera pose is equal to compute the maximum posterior distribution $P(\mathbb{T} \mid \mathbb{X} ; \boldsymbol{\theta}, \mathbb{W})$, which is not tractable since it requires to integral over $\mathbb{T}$ to compute $P(\mathbb{X})$. Thus, we use a Monte-Carlo based approximation, where for each depth map position θ^{j_1} we randomly sample two additional distinct positions $\{j_2, j_3\}$ to form a 3-tuple $\boldsymbol{\Theta^g} = \{\theta^{j_1}, \theta^{j_2}, \theta^{j_3}\}$, with associated rigidness values $\mathbb{W}^g_t = \{W^{j_1}_t, W^{j_2}_t, W^{j_3}_t\}$, to represent the g^{th} group. Then the posterior can be approximated as

$$\begin{aligned} & P(\mathbb{T} \mid \mathbb{X} ; \boldsymbol{\theta}, \mathbb{W}) \\ & \approx \prod_t \left[\frac{1}{S} \sum_g^S P(\boldsymbol{T_t} \mid \boldsymbol{X}_{\boldsymbol{t}}^{\boldsymbol{\pi_{t-1}}(\boldsymbol{\Theta^g})} ; \boldsymbol{\Theta^g}, \mathbb{W}^g_t) \right] \end{aligned} \tag{16}$$

where S is the total number of groups. Although the posterior $P(\boldsymbol{T_t} \mid \boldsymbol{X}_{\boldsymbol{t}}^{\boldsymbol{\pi_{t-1}}(\boldsymbol{\Theta^g})} ; \boldsymbol{\Theta^g}, \mathbb{W}^g_t)$ is still not tractable, using 3 pairs of 3D-2D correspondences, PnP reaches its minimal form of P3P, which can be solved efficiently using the P3P algorithm [24, 48, 55, 42]. Hence we have

$$\hat{\boldsymbol{T}}^{\boldsymbol{g}}_{\boldsymbol{t}} = \phi(\left(\prod_{i=0}^{t-1} \boldsymbol{T_i}\right) \boldsymbol{\Theta^g} \boldsymbol{K^{-1}} [x_g \; y_g \; 1]^T, \boldsymbol{\pi_{t-1}}(\boldsymbol{\Theta^g}) + \boldsymbol{X}_{\boldsymbol{t}}^{\boldsymbol{\pi_{t-1}}(\boldsymbol{\Theta^g})}) \tag{17}$$

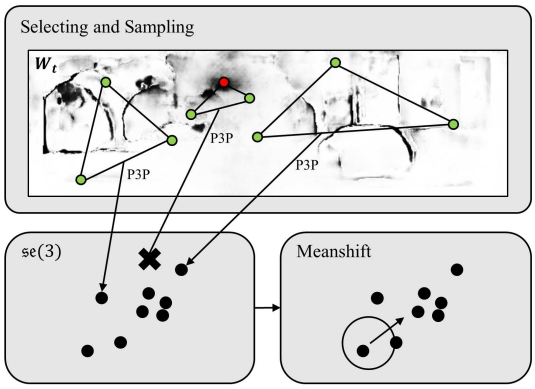


Figure 6: **Pose MAP approximation via meanshift-based mode search.** Each 3D-2D correspondence is part of a unique minimal P3P instance, constituting a pose sample that is weighted by the rigidness map. We map samples to $\mathfrak{se}(3)$ and run meanshift to find the mode.

where $\phi(\cdot,\cdot)$ denotes the P3P solver, for which we use AP3P [42]. The first input argument indicates the 3D coordinates of selected depth map pixels at time $t-1$ obtained by combining previous camera poses, while the second input argument is their 2D correspondences at time t, obtained using optical flow displacement. Hence, we use a tractable variational distribution $q(\boldsymbol{T}_t^{\boldsymbol{g}})$ to approximate its true posterior.

$$q(\boldsymbol{T}_t^{\boldsymbol{g}}) \sim \mathcal{N}(\hat{\boldsymbol{T}}_t^{\boldsymbol{g}}, \boldsymbol{\Sigma}) \tag{18}$$

where $q(\boldsymbol{T}_t^{\boldsymbol{g}})$ is a normal distribution with mean $\hat{\boldsymbol{T}}_t^{\boldsymbol{g}}$ and predefined fixed covariance matrix $\boldsymbol{\Sigma}$ for simplicity. Furthermore, we weight each variational distribution with $\|\mathbb{W}_t^g\| = \prod_{W_i \in \mathbb{W}_t^g} W_i$, such that potential outliers indicated by rigidness maps can be excluded or down weighted. Then, the full posterior can be approximated as

$$P(\mathbb{T} \mid \mathbb{X}; \boldsymbol{\theta}, \mathbb{W}) \approx \prod_t \left[\frac{1}{\sum_g \|\mathbb{W}_t^g\|} \sum_g \|\mathbb{W}_t^g\| q(\boldsymbol{T}_t^{\boldsymbol{g}}) \right] \tag{19}$$

We have approximated the posterior $P(\mathbb{T} \mid \mathbb{X}; \boldsymbol{\theta}, \mathbb{W})$ with a weighted combination of $q(\boldsymbol{T}_t^{\boldsymbol{g}})$. Solving for the optimum $\boldsymbol{T}_t^*$ on the posterior equates to finding the mode of the posterior. Since we assume all $q(\boldsymbol{T}_t^{\boldsymbol{g}})$ to share the same covariance structure, mode finding on this distribution equates to applying meanshift [10] with a Gaussian kernel of covariance $\boldsymbol{\Sigma}$. Note that since $\hat{\boldsymbol{T}}_t^{\boldsymbol{g}}$ lies in $SE(3)$ [4], while meanshift is applied to vector space, an obtained a mode can not be guaranteed to lie in $SE(3)$. Thus, poses $\hat{\boldsymbol{T}}_t^{\boldsymbol{g}}$ are first converted to a 6-vector $\boldsymbol{p} = log(\hat{\boldsymbol{T}}_t^{\boldsymbol{g}}) \in \mathfrak{se}(3)$ in Lie algebra through logarithm mapping, and meanshift is applied to the 6-vector space.

Input: Optical flow sequence $\mathbb{X} = \{\boldsymbol{X_t} \mid t = 1 \cdots t_N\}$
Output: Camera poses $\mathbb{T} = \{\boldsymbol{T_t} \mid t = 1 \cdots t_N\}$ Depth map $\boldsymbol{\theta}$ of the first frame
Initialize $\mathbb{W} = \{\boldsymbol{W_t} \mid t = 1 \cdots t_N\}$ all to one
Initialize $\boldsymbol{T_1}$ using epipolar geometry from $\boldsymbol{X_1}$
Triangulate $\boldsymbol{\theta}$ from $\boldsymbol{T_1}$ and $\boldsymbol{X_1}$
Repeat until $\mathbb{T}$ converges
For $i = 1 \cdots t$
Update $\boldsymbol{T_i}$ according to Eq. (19)
Update and smooth $\mathbb{W}$ according to Eq. (12)
Update $\boldsymbol{\theta}$ according to Eq. (10)
Update $\mathbb{W}$ according to Eq. (12) w/o smoothing

Table 1: **The VOLDOR algorithm.**

5.3. Algorithm Integration

We now describe the integrated workflow of our visual odometry algorithm, which we denote VOLDOR. Per Table 1, our input is a sequence of dense optical flows $\mathbb{X} = \{\boldsymbol{X_t} \mid t = 1 \cdots t_N\}$, and our output will be the camera poses of each frame $\mathbb{T} = \{\boldsymbol{T_t} \mid t = 1 \cdots t_N\}$ as well as the depth map $\boldsymbol{\theta}$ of the first frame. Usually, 4-8 optical flows per batch are used. Firstly, VOLDOR initializes all $\mathbb{W}$ to one and $\boldsymbol{T_1}$ is initialized from epipolar geometry estimated from $\boldsymbol{X_1}$ using a least median of squares estimator [68] or, alternatively, from previous estimates if available (i.e. overlapping consecutive frame batches). Then, $\boldsymbol{\theta}$ is obtained from two-view triangulation using $\boldsymbol{T_1}$ and $\boldsymbol{X_1}$. Next, the optimization loop between camera pose, depth map and rigidness map runs until convergence, usually within 3 to 5 iterations. Note we did not smooth the rigidness map before updating camera poses to prevent loss of fine details indicating the potential high-frequency noise in the observations.

6. Experiments

6.1. KITTI Benchmark

We tested on the KITTI odometry benchmark [25] of a car driving at urban and highway environments. We use PWC-Net [77] as our external dense optical flow input. The sliding window size is set to 6 frames. We set λ in Eq. (8) to 0.15, γ in Eq. (11) to 0.9. The Gaussian kernel covariance matrix $\boldsymbol{\Sigma}$ in Eq. (18) is set to diagonal, scaled to 0.1 and 0.004 at the dimensions of translation and rotation respectively. The hyper-parameters for the Fisk residual model are $a_1 = 0.01$, $a_2 = 0.09$, $b_1 = -0.0022$, $b_2 = 1.0$, obtained from Figure 3. Finally, absolute scale is estimated from ground plane by taking the mode of pixels with surface normal vector near perpendicular. More details of ground plane estimation are provided in the appendix.

Table 2 and Figure 7 are our results on the KITTI odometry training set sequences 0-10. We picked VISO2 [26]

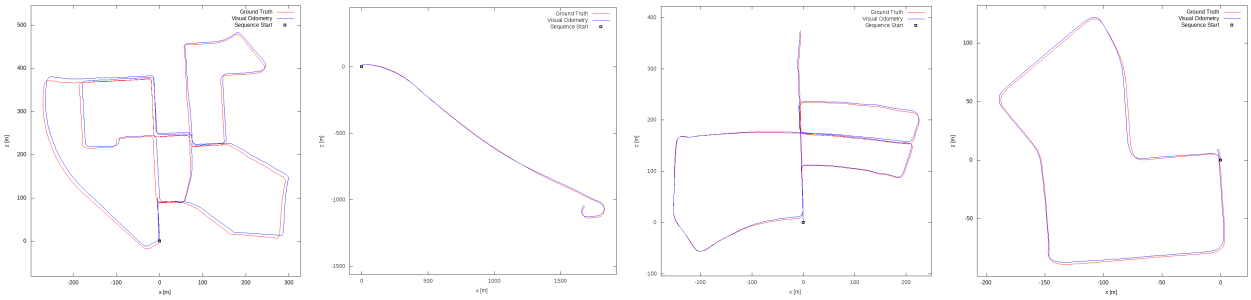


Figure 7: **Results on KITTI sequences 00, 01, 05, 07**. Monocular VOLDOR does not deploy any mapping, bundle adjustment or loop closure. Scale is estimated assuming fixed and known camera height from the ground.

Sequence	VISO2-M		MLM-SFM		VOLDOR	
	Trans. (%)	Rot. (deg/m)	Trans. (%)	Rot. (deg/m)	Trans. (%)	Rot. (deg/m)
00	12.53	0.0260	2.04	0.0048	**1.09**	**0.0039**
01	28.09	0.0641	-	-	**2.31**	**0.0037**
02	3.98	0.0123	1.50	**0.0035**	**1.19**	0.0042
03	4.09	0.0206	3.37	**0.0021**	**1.46**	0.0034
04	2.58	0.0162	1.43	**0.0023**	**1.13**	0.0049
05	14.68	0.0379	2.19	**0.0038**	**1.15**	0.0041
06	6.73	0.0195	2.09	0.0081	**1.13**	**0.0045**
07	14.95	0.0558	-	-	**1.63**	**0.0054**
08	11.63	0.0215	2.37	**0.0044**	**1.50**	**0.0044**
09	4.94	0.0140	1.76	0.0047	**1.61**	**0.0039**
10	23.36	0.0348	2.12	0.0085	**1.44**	**0.0043**
Avg.	10.85	0.0249	2.03	0.0045	**1.32**	**0.0042**

Table 2: **Results on KITTI training sequences 0-10**. The translation and rotation errors are averaged over all sub-sequences of length from 100 meters to 800 meters with 100 meter steps.

Method	Trans. (%)	Rot. (deg/m)
VISO2-M [26]	11.94	0.0234
MLM-SFM [74] [75]	2.54	0.0057
PbT-M2 [19] [17] [18]	2.05	0.0051
BVO [62]	1.76	**0.0036**
VOLDOR (Ours)	**1.65**	0.0050
VO3pt* [2] [1]	2.69	0.0068
VISO2-S* [26]	2.44	0.0114
eVO* [69]	1.76	0.0036
S-PTAM* [63] [64]	1.19	0.0025
ORB-SLAM2* [57]	1.15	0.0027

Table 3: **Results on KITTI odometry testing sequences 11-21.** * indicates the method is based on stereo input.

and MLM-SFM [74] [75] as our baselines, whose scales are also estimated from ground height. Table 3 compares our result with recent popular methods on KITTI test set sequences 11-21, where we download from KITTI odometry official ranking board. As the results shows, VOLDOR has achieved top-ranking accuracy under KITTI dataset among monocular methods.

Table 4 shows our depth map quality on the KITTI stereo benchmark. We masked out foreground moving object as well as aligned our depth with groundtruth to solve scale ambiguity. We separately evaluated the depth quality for

Methods	Density	EPE / px	bg-outlier
GC-Net [43]	**100.00%**	0.7	2.21%
PSMNet [8]	**100.00%**	0.6	1.86%
GA-Net-15 [95]	**100.00%**	**0.5**	**1.55%**
Ours ($W^j > 5$)	27.07%	**0.5616**	**1.47%**
Ours ($W^j > 4$)	37.87%	0.6711	2.05%
Ours ($W^j > 3$)	49.55%	0.7342	2.56%
Ours ($W^j > 2$)	62.50%	0.8135	3.17%
Ours ($W^j > 1$)	78.18%	0.9274	4.17%
Ours ($W^j > 0$)	**100.00%**	1.2304	5.82%

Table 4: **Results on KITTI stereo benchmark.** A pixel is considered as outlier if disparity EPE is $> 3px$ and $> 5\%$. W^j denotes the sum of pixel rigidness $W^j = \sum_t W_t^j$.

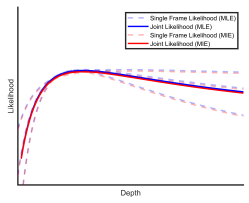
Sequence	ORB-SLAM2 (RGB-D)	DVO-SLAM (RGB-D)	DSO (Mono)	Ours (Mono)
fr1/desk	0.0163	0.0185	0.0168	**0.0133**
fr1/desk2	0.0162	0.0238	0.0188	**0.0150**
fr1/room	0.0102	0.0117	0.0108	**0.0090**
fr2/desk	**0.0045**	0.0068	0.0048	0.0053
fr2/xyz	0.0034	0.0055	**0.0025**	0.0034
fr3/office	0.0046	0.0102	0.0050	**0.0045**
fr3/nst	0.0079	0.0073	0.0087	**0.0071**

Table 5: **Results on TUM RGB-D dataset.** The values are translation RMSE in meters.

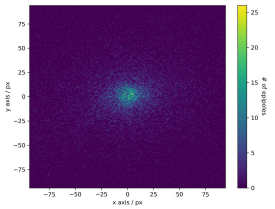
different rigidness probabilities, where $W^j = \sum_t W_t^j$. The EPE of PSMNet [8], GC-Net [43] and GA-Net [95] are measured on stereo 2012 test set and background outlier percentage is measured on stereo 2015 test set while our method is measured on training set on stereo 2015.

6.2. TUM RGB-D Benchmark

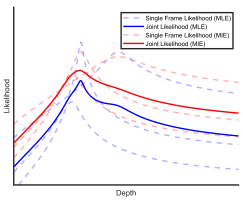
Accuracy experiments on TUM RGB-D [76] compared VOLDOR vs. full SLAM systems. In all instances, we rigidly align trajectories to groundtruth for segments with 6 frames and estimate mean translation RMSE of all segments. Parameters remains the same to KITTI experiments. Our comparative baselines are an indirect sparse method ORB-SLAM2 [57], a direct sparse method DSO [15] and a dense direct method DVO-SLAM [45]. Per Table 5,



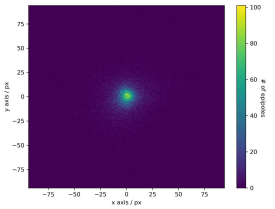
(a) Depth likelihood with Gaussian-(MLE/MIE) Residual Model



(b) Epipole distribution with Gaussian Residual Model

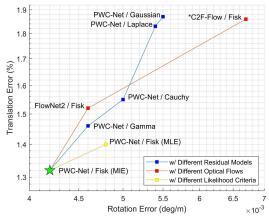


(c) Depth likelihood with Fisk-(MLE/MIE) Residual Model

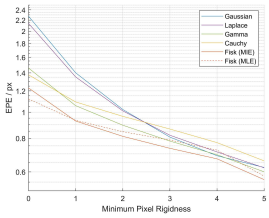


(d) Epipole distribution with Fisk Residual Model

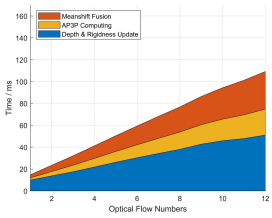
Figure 8: **Fisk model qualitative study.** (a) and (c) visualize the depth likelihood function under Gaussian and Fisk residual models with MLE and MIE criteria. Dashed lines indicate the likelihood given by a single optical flow. Solid lines are joint likelihood obtained by fusing all dashed lines. MLE and MIE are shown in different colors. (c) and (d) visualize the epipole distribution for 40K camera pose samples. For better visualization, the density color bars of (b) (d) are scaled differently.



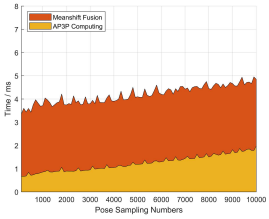
(a) Ablation study over camera pose



(b) Ablation study over depth



(c) Timing over frame numbers



(d) Timing over pose samples

Figure 9: **Ablation study and runtime.** (a) shows the camera pose error of VOLDOR under different residual models and dense optical flow input (*due to noisy ground estimations given by C2F-Flow, its scale is corrected using groundtruth). (b) shows our depth map accuracy under different residual models. (c) and (d) show the runtime of our method tested on a GTX 1080Ti GPU.

VOLDOR performs well under indoor capture exhibiting smaller camera motions and diverse motion patterns.

6.3. Ablation and Performance Study

Figure 8 visualizes the depth likelihood function and camera pose sampling distribution. With our Fisk residual model, depth likelihood from each single frame has a well localized extremum (Fig.8-c), compared to a Gaussian residual model (Fig.8-a). This leads to a joint likelihood having a more distinguishable optimum, and results in more concentrated camera pose samplings (Fig.8-b,d). Also, with the MIE criteria, depth likelihood of the Fisk residual model is relaxed to a smoother shape whose effectiveness is further analyzed in Fig. 9, while a Gaussian residual model is agnostic to the choice between MIE and MLE (Fig.8-a). Per the quantitative study in Fig. 9 (b), compared to other analytic distributions, the Fisk residual model gives significantly better depth estimation when only a small number of reliable observations (low W^j) are used. Performance across different residual models tends to converge as the number of reliable samples increases (high W^j), while the Fisk residual model still provides the lowest EPE. Figure 9 (a) shows the ablation study on camera pose estimation for three optical flow methods, four residual models and our proposed MIE criteria. The accuracy of combining PWC-Net optical flow, a Fisk residual model, and our MIE criteria, strictly dominates (in the Pareto sense) all other combinations. Figure 9 (b) shows the MIE criteria yields depth estimates that are more consistent across the entire image sequence, leading to improved overall accuracy. However, at the extreme case of predominantly unreliable observations (very low W^j) MLE provides the most accurate depth. Figure 9 (c) shows the overall runtime evaluation for each component under different frame numbers. Figure 9 (d) shows runtime for pose update under different sample rates.

7. Conclusion

Conceptually, we pose the VO problem as an instance of geometric parameter inference under the supervision of an adaptive model of the empirical distribution of dense optical flow residuals. Pragmatically, we develop a monocular VO pipeline which obviates the need for a) feature extraction, b) ransac-based estimation, and c) local bundle adjustment, yet still achieves top-ranked performance in the KITTI and TUM RGB-D benchmarks. We posit the use of dense-indirect representations and adaptive data-driven supervision as a general and extensible framework for multi-view geometric analysis tasks.

References

[1] P Fernández Alcantarilla. Vision based localization: from humanoid robots to visually impaired people. *Electronics (University of Alcala, 2011)*, 2011. 7

[2] Pablo F Alcantarilla, José J Yebes, Javier Almazán, and Luis M Bergasa. On combining visual slam and dense scene flow to increase the robustness of localization and mapping in dynamic environments. In *2012 IEEE International Conference on Robotics and Automation*, pages 1290–1297. IEEE, 2012. 7

[3] G Jogesh Babu and Calyampudi R Rao. Goodness-of-fit tests when parameters are estimated. *Sankhya*, 66(1):63–74, 2004. 4

[4] Jose-Luis Blanco. A tutorial on se (3) transformation parameterizations and on-manifold optimization. *University of Malaga, Tech. Rep*, 3, 2010. 6

[5] Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J Davison. Codeslamlearning a compact, optimisable representation for dense visual slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2560–2568, 2018. 2

[6] Behzad Bozorgtabar, Mohammad Saeed Rad, Dwarikanath Mahapatra, and Jean-Philippe Thiran. Syndemo: Synergistic deep feature alignment for joint learning of depth and ego-motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4210–4219, 2019. 2

[7] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012. 4

[8] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 7

[9] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7063–7072, 2019. 2

[10] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799, 1995. 6

[11] Ronald Clark, Michael Bloesch, Jan Czarnowski, Stefan Leutenegger, and Andrew J Davison. Ls-net: Learning to solve nonlinear least squares for monocular stereo. *arXiv preprint arXiv:1809.02966*, 2018. 2

[12] Gabriele Costante, Michele Mancini, Paolo Valigi, and Thomas A Ciarfuglia. Exploring representation learning with cnns for frame-to-frame ego-motion estimation. *IEEE robotics and automation letters*, 1(1):18–25, 2015. 2

[13] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):1052–1067, 2007. 2

[14] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 2

[15] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. 2, 7

[16] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014. 1, 2

[17] Nolang Fanani, Matthias Ochs, Henry Bradler, and Rudolf Mester. Keypoint trajectory estimation using propagation based tracking. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 933–939. IEEE, 2016. 7

[18] Nolang Fanani, Alina Stürck, Marc Barnada, and Rudolf Mester. Multimodal scale estimation for monocular visual odometry. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1714–1721. IEEE, 2017. 7

[19] Nolang Fanani, Alina Stürck, Matthias Ochs, Henry Bradler, and Rudolf Mester. Predictive monocular odometry (pmo): What is possible without ransac and multiframe bundle adjustment? *Image and Vision Computing*, 68:3–13, 2017. 7

[20] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 15–22. IEEE, 2014. 2

[21] Friedrich Fraundorfer and Davide Scaramuzza. Visual odometry: Part i: The first 30 years and fundamentals. *IEEE Robotics and Automation Magazine*, 18(4):80–92, 2011. 1

[22] Friedrich Fraundorfer and Davide Scaramuzza. Visual odometry: Part ii: Matching, robustness, optimization, and applications. *IEEE Robotics & Automation Magazine*, 19(2):78–90, 2012. 1

[23] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018. 2

[24] Xiao-Shan Gao, Xiao-Rong Hou, Jianliang Tang, and Hang-Fei Cheng. Complete solution classification for the perspective-three-point problem. *IEEE transactions on pattern analysis and machine intelligence*, 25(8):930–943, 2003. 5

[25] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 2, 4, 6

[26] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 963–968. Ieee, 2011. 6, 7

[27] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE International*

Conference on Computer Vision, pages 3828–3838, 2019. 2

[28] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017. 2

[29] Ruben Gomez-Ojeda and Javier Gonzalez-Jimenez. Robust stereo visual odometry through a probabilistic combination of points and line segments. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2521–2526. IEEE, 2016. 2

[30] Ruben Gomez-Ojeda, Francisco-Angel Moreno, and Javier Gonzalez-Jimenez. Accurate stereo visual odometry with gamma distributions. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1423–1428. IEEE, 2017. 3

[31] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 484–500, 2018. 2

[32] Jiahui Huang, Sheng Yang, Zishuo Zhao, Yu-Kun Lai, and Shi-Min Hu. Clusterslam: A slam backend for simultaneous rigid body clustering and motion estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5875–5884, 2019. 2

[33] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8981–8989, 2018. 2

[34] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2019. 2

[35] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 1, 2, 3

[36] Eddy Ilg, Tonmoy Saikia, Margret Keuper, and Thomas Brox. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 614–630, 2018. 2

[37] Andrew Jaegle, Stephen Phillips, and Kostas Daniilidis. Fast, robust, continuous monocular egomotion computation. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 773–780. IEEE, 2016. 3

[38] Joel Janai, Fatma Guney, Anurag Ranjan, Michael Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 690–706, 2018. 2

[39] Kenichi Kanatani. Statistical optimization for geometric fitting: Theoretical accuracy bound and high order error analysis. *International Journal of Computer Vision*, 80(2):167–188, 2008. 1

[40] Ken-ichi Kanatani. Uncertainty modeling and model selection for geometric inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1307–1319, 2004. 1

[41] Yasushi Kanazawa and Kenichi Kanatani. Do we really have to consider covariance matrices for image feature points? *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, 86(1):1–10, 2003. 1

[42] Tong Ke and Stergios I Roumeliotis. An efficient algebraic solution to the perspective-three-point problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7225–7233, 2017. 5, 6

[43] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017. 7

[44] Christian Kerl, Jorg Stuckler, and Daniel Cremers. Dense continuous-time tracking and mapping with rolling shutter rgb-d cameras. In *Proceedings of the IEEE international conference on computer vision*, pages 2264–2272, 2015. 2

[45] Christian Kerl, Jürgen Sturm, and Daniel Cremers. Dense visual slam for rgb-d cameras. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2100–2106. IEEE, 2013. 1, 2, 7

[46] Pyojin Kim, Brian Coltin, and H Jin Kim. Linear rgb-d slam for planar environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 333–348, 2018. 2

[47] Bernd Kitt, Andreas Geiger, and Henning Lategahn. Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme. In *2010 ieee intelligent vehicles symposium*, pages 486–492. IEEE, 2010. 1, 2

[48] Laurent Kneip, Davide Scaramuzza, and Roland Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *CVPR 2011*, pages 2969–2976. IEEE, 2011. 5

[49] Shunkai Li, Fei Xue, Xin Wang, Zike Yan, and Hongbin Zha. Sequential adversarial learning for self-supervised deep visual odometry. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2851–2860, 2019. 2

[50] Ce Liu et al. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009. 3

[51] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4571–4580, 2019. 2

[52] Shing Yan Loo, Ali Jahani Amiri, Syamsiah Mashohor, Sai Hong Tang, and Hong Zhang. Cnn-svo: Improving the mapping in semi-direct visual odometry using single-image depth prediction. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5218–5223. IEEE, 2019. 2

[53] J. Ma, J. Zhao, J. Tian, A. L. Yuille, and Z. Tu. Robust point matching via vector field consensus. *IEEE Transactions on Image Processing*, 23(4):1706–1721, April 2014. 4

[54] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, 2018. 2

[55] Andreas Masselli and Andreas Zell. A new geometric approach for faster solving the perspective-three-point problem. In *2014 22nd International Conference on Pattern Recognition*, pages 2119–2124. IEEE, 2014. 5

[56] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 2

[57] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. 1, 2, 7

[58] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. Dtam: Dense tracking and mapping in real-time. In *2011 International Conference on Computer Vision*, pages 2320–2327, Nov 2011. 1, 2

[59] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, 26(6):0756–777, 2004. 2

[60] David Nistér, Oleg Naroditsky, and James Bergen. Visual odometry for ground vehicle applications. *Journal of Field Robotics*, 23(1):3–20, 2006. 1

[61] Geoffrey Pascoe, Will Maddern, Michael Tanner, Pedro Piniés, and Paul Newman. Nid-slam: Robust monocular slam using normalised information distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1435–1444, 2017. 2

[62] Fabio Pereira, Joel Luft, Gustavo Ilha, Arthur Sofiatti, and Altamiro Susin. Backward motion for estimation enhancement in sparse visual odometry. In *2017 Workshop of Computer Vision (WVC)*, pages 61–66. IEEE, 2017. 7

[63] Taihú Pire, Thomas Fischer, Gastón Castro, Pablo De Cristóforis, Javier Civera, and Julio Jacobo Berlles. S-ptam: Stereo parallel tracking and mapping. *Robotics and Autonomous Systems*, 93:27–42, 2017. 7

[64] Taihú Pire, Thomas Fischer, Javier Civera, Pablo De Cristóforis, and Julio Jacobo Berlles. Stereo parallel tracking and mapping for robot localization. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1373–1378. IEEE, 2015. 1, 2, 7

[65] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4161–4170, 2017. 2

[66] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12240–12249, 2019. 2

[67] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1164–1172, 2015. 3

[68] Peter J Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984. 6

[69] Martial Sanfourche, Vincent Vittori, and Guy Le Besnerais. evo: A realtime embedded stereo odometry for mav applications. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2107–2114. IEEE, 2013. 7

[70] Thomas Schops, Torsten Sattler, and Marc Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 134–144, 2019. 2

[71] David Schubert, Nikolaus Demmel, Vladyslav Usenko, Jorg Stuckler, and Daniel Cremers. Direct sparse odometry with rolling shutter. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 682–697, 2018. 2

[72] Lu Sheng, Dan Xu, Wanli Ouyang, and Xiaogang Wang. Unsupervised collaborative learning of keyframe detection and visual odometry towards monocular deep slam. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4302–4311, 2019. 2

[73] S. Song and M. Chandraker. Robust scale estimation in real-time monocular sfm for autonomous driving. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1566–1573, June 2014. 4

[74] Shiyu Song and Manmohan Chandraker. Robust scale estimation in real-time monocular sfm for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1566–1573, 2014. 7

[75] Shiyu Song, Manmohan Chandraker, and Clark C Guest. Parallel, real-time monocular visual odometry. In *2013 ieee international conference on robotics and automation*, pages 4698–4705. IEEE, 2013. 7

[76] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 573–580. IEEE, 2012. 2, 7

[77] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018. 1, 2, 3, 6

[78] Chengzhou Tang and Ping Tan. Ba-net: Dense bundle adjustment network. *arXiv preprint arXiv:1806.04807*, 2018. 2

[79] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6243–6252, 2017. 2

[80] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5038–5047, 2017. 2

[81] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, 2018. 2

[82] Rui Wang, Stephen M Pizer, and Jan-Michael Frahm. Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5555–5564, 2019. 2

[83] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2043–2050. IEEE, 2017. 2

[84] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8071–8081, 2019. 2

[85] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4884–4893, 2018. 2

[86] Thomas Whelan, Stefan Leutenegger, R Salas-Moreno, Ben Glocker, and Andrew Davison. Elasticfusion: Dense slam without a pose graph. Robotics: Science and Systems, 2015. 2

[87] Jonas Wulff, Laura Sevilla-Lara, and Michael J Black. Optical flow in mostly rigid scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4671–4680, 2017. 2

[88] Fei Xue, Xin Wang, Shunkai Li, Qiuyuan Wang, Junqiu Wang, and Hongbin Zha. Beyond tracking: Selecting memory and refining poses for deep visual odometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8575–8583, 2019. 2

[89] Nan Yang, Rui Wang, Jorg Stuckler, and Daniel Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 817–833, 2018. 2

[90] Yanchao Yang and Stefano Soatto. Conditional prior networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 271–287, 2018. 2

[91] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1983–1992, 2018. 2

[92] Masashi Yokozuka, Shuji Oishi, Simon Thompson, and Atsuhiko Banno. Vitamin-e: Visual tracking and mapping with extremely dense feature points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9641–9650, 2019. 2

[93] Bernhard Zeisl, Pierre Fite Georgel, Florian Schweiger, Eckehard G Steinbach, Nassir Navab, and G Munich. Estimation of location uncertainty for scale invariant features points. In *BMVC*, pages 1–12, 2009. 1

[94] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 340–349, 2018. 2

[95] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip H.S. Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 7

[96] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4106–4115, 2019. 2

[97] Cheng Zhao, Li Sun, Pulak Purkait, Tom Duckett, and Rustam Stolkin. Learning monocular visual odometry with dense 3d mapping from dense 3d flow. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6864–6871. IEEE, 2018. 2

[98] Yipu Zhao and Patricio A Vela. Good line cutting: Towards accurate pose tracking of line-assisted vo/vslam. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 516–531, 2018. 2

[99] Enliang Zheng, Enrique Dunn, Vladimir Jojic, and Jan-Michael Frahm. Patchmatch based joint view selection and depthmap estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1517, 2014. 4

[100] Yinqiang Zheng, Yubin Kuang, Shigeki Sugimoto, Kalle Astrom, and Masatoshi Okutomi. Revisiting the pnp problem: A fast, general and optimal solution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2344–2351, 2013. 5

[101] Yiran Zhong, Pan Ji, Jianyuan Wang, Yuchao Dai, and Hongdong Li. Unsupervised deep epipolar flow for stationary or dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12095–12104, 2019. 2

[102] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017. 2

[103] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 36–53, 2018. 2