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 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.

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-
posses 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-geometric 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 \( \mathbf{X} = \{ \mathbf{X}_t \mid t = 1, \cdots, t_N \} \), where \( \mathbf{X}_t \) is the optical flow map from image \( \mathbf{I}_{t-1} \) to \( \mathbf{I}_t \), while \( \mathbf{X}_t^j = (u_j^t, v_j^t)^T \) denotes the optical flow vector of pixel \( j \) at time \( t \). We aim to infer the camera poses \( \mathcal{T} = \{ \mathbf{T}_t \mid t = 1, \cdots, t_N \} \), where \( \mathbf{T}_t \in SE(3) \) represents the relative motion from time \( t - 1 \) to \( t \).

To define a likelihood model relating our observations \( \mathbf{X} \)
to T, we introduce two additional (latent) variable types: 1) a depth field \( \theta \) defined over \( I_0 \); where we denote \( \theta^j \) as the depth value at pixel \( j \), and 2) a rigidity probability map \( \mathcal{W}_t \) associated to \( \theta \) at time \( t \); while \( \mathcal{W} = \{ \mathcal{W}_t \mid t = 1, \cdots, t_N \} \) denotes the set of rigidity maps, and \( \mathcal{W}_t^j \) denotes the rigidity probability of pixel \( j \) at time \( t \).

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

\[
\pi_t(\theta^j) = K \left( \prod_{i=0}^{t-1} T_i \right) \theta^j K^{-1} \begin{bmatrix} x_j \\ y_j \\ 1 \end{bmatrix}
\]

(1)

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

**Mixture Likelihood Model** We model the residuals between observation flow and rigid flow with respect to the continuous rigidity probability \( \mathcal{W}_t \).

\[
P(X_t^{\pi_{t-1}(\theta^j)} \mid \theta^j, T_t, \mathcal{W}_t; T_1 \cdots T_{t-1}) = \begin{cases} 
\rho(\xi_t(\theta^j) \mid X_t^{\pi_{t-1}(\theta^j)}) & \text{if } \mathcal{W}_t^j = 1 \\
\mu(X_t^{\pi_{t-1}(\theta^j)}) & \text{if } \mathcal{W}_t^j = 0 
\end{cases}
\]

(2)

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

At this point, our visual odometry problem can be modeled as a maximum likelihood estimation problem

\[
\arg\max_{\theta, T, \mathcal{W}} P(X_0, T, \mathcal{W}) = \arg\max_{\theta, T, \mathcal{W}} \prod_{t=0}^{T_N} P(X_t^{\pi_{t-1}(\theta^j)} \mid \theta^j, T_t, \mathcal{W}_t) \quad (3)
\]

Furthermore, we promote spatial consistency among both our dense hidden variables \( \theta \) and \( \mathcal{W} \), 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.

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.

Here, given \( v_{ob} = X^{	heta}_{t} P^{a-1}(\theta^i) \), we model the probability of \( v_{rig} = \xi_t(\theta^i) \) matching the underlying groundtruth, as

\[
\rho(v_{rig} | v_{ob}) = \mathcal{F}(\lVert v_{rig} - v_{ob} \rVert_2^2; A(v_{ob}), B(v_{ob})) \tag{4}
\]

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

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

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

\[
A(v_{ob}) = a_1 e^{a_2 \|v_{ob}\|^2} \tag{6}
\]

\[
B(v_{ob}) = b_1 \|v_{ob}\|^2 + b_2 \tag{7}
\]

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,

\[
\mu(v_{ob}) = \mathcal{F}(\lambda^2 \|v_{ob}\|^2; A(v_{ob}), B(v_{ob})) \tag{8}
\]

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

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) \( \alpha, \beta \) are estimated from KITTI empirical data using, respectively, log-linear and linear regression, as described in Eq. (6), (7).

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 \( \theta \) and its rigidness \( W \) over time while assuming fixed known camera pose \( T \). We approximate the true posterior \( P(X | \theta, W; T) \) through a GEM framework [72]. In this section, we will denote Eq. (2) as \( P(X | \theta, W) \), where the fixed \( T \) is omitted. We approximate the intractable real posterior \( P(\theta, W | X) \) with a restricted family of distributions \( q(\theta, W) \), where \( q(\theta, W) = \prod_t q(\theta^t) \prod_t q(W_t) \).

For tractability, \( q(\theta^t) \) is further constrained to the family of Kronecker delta functions \( q(\theta^t) = \delta(\theta^t = \theta^*, \theta^*) \), where \( \theta^* \) is a parameter to be estimated. Moreover, \( q(W_t) \) inherits the smoothness defined on the rigidness map \( W_t \) as minimizing KL divergence between the variational distribution and the true posterior. In the M step, we seek to estimate an optimal value for \( \theta^* \) given an estimated PDF on \( W_t \). Next, we describe our choice for the estimators used for this task.

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

\[
\hat{\theta}_{MLE} = \arg\max_{\theta^*} \sum_t q(W_t) \log P(X_t^\theta P(a-1) | \theta^t = \theta^*, W_t^i) \tag{9}
\]

where \( q(W_t) \) 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
the rigidity probabilities $\mathcal{W}$, 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,

$$\theta_{\text{MIE}}^{j} = \arg\max_{\theta^{j}} \sum_{t} q(W_{t}^{j} = 1) \log \frac{P(X_{t-1}^{\theta^{j}})P(W_{t}^{j} = 1)}{\sum_{W_{t}^{j}} P(X_{t-1}^{\theta^{j}})P(W_{t}^{j} = 1)}$$

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

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

**Updating the Rigidity 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 rigidity map

$$P(W_{t}^{j} | W_{t-1}^{j-1}) = \begin{pmatrix} \gamma & 1 - \gamma \\ 1 - \gamma & \gamma \end{pmatrix}$$

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

$$q(W_{t}^{j}) = \frac{1}{A} m_f(W_{t}^{j})m_b(W_{t}^{j})$$

where $A$ is a normalization factor while $m_f(W_{t}^{j})$ and $m_b(W_{t}^{j})$ are the forward and backward message of $W_{t}^{j}$ computed recursively as

$$m_f(W_{t}^{j}) = P_{\text{ems}}^{j} \sum_{W_{t-1}^{j-1}} m_f(W_{t}^{j-1})P(W_{t}^{j} | W_{t-1}^{j-1})$$

$$m_b(W_{t}^{j}) = \sum_{W_{t+1}^{j+1}} m_b(W_{t}^{j+1})P_{\text{ems}}^{j+1} P(W_{t}^{j+1} | W_{t}^{j})$$

where $P_{\text{ems}}^{j}$ is the emission probability refer to Eq. (2), $P_{\text{ems}}^{j} = P(X_{t}^{\theta_{t-1}^{j}} | \theta_{t}^{j}, W_{t}^{j}; T)$.

**5.2. Pose Update**

We update camera poses while assuming fixed known depth $\theta$ and rigidity maps $\mathcal{W}$. We use the optical flow chains in $\mathcal{X}$ 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 rigidity map (all one).

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

$$\arg\max_{\pi} P(\pi | \mathcal{X} ; \theta, \mathcal{W})$$

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

$$P(\pi | \mathcal{X} ; \theta, \mathcal{W}) \approx \prod_{t} \left[ \frac{1}{S} \sum_{g} P(T_{t} | X_{t}^{\pi_{t-1}^{g}(\Theta^{g})} ; \Theta^{g}, \mathcal{W}_{g}^{t}) \right]$$

where $S$ is the total number of groups. Although the posterior $P(T_{t} | X_{t}^{\pi_{t-1}^{g}(\Theta^{g})} ; \Theta^{g}, \mathcal{W}_{g}^{t})$ 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

$$\tilde{T}_{g}^{t} = \phi \left( \prod_{i=0}^{t-1} T_{i} \right) \Theta^{g} K^{-1} [x_{g} y_{g} 1]^T, \pi_{t-1}^{g}(\Theta^{g}) + X_{t}^{\pi_{t-1}^{g}(\Theta^{g})}, \pi_{t-1}^{g}(\Theta^{g}) + X_{t}^{\pi_{t-1}^{g}(\Theta^{g})}$$


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 \( q(T^q_t) \) to approximate its true posterior.

\[
q(T^q_t) \sim \mathcal{N}(\hat{T}^q_t, \Sigma) \tag{18}
\]

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

\[
P(T \mid X; \theta, W) \approx \prod_i \frac{1}{\sum_g ||W^q_t||} \sum_g ||W^q_t|| q(T^q_t) \tag{19}
\]

We have approximated the posterior \( P(T \mid X; \theta, W) \) with a weighted combination of \( q(T^q_t) \). Solving for the optimum \( T^q_t \) on the posterior equates to finding the mode of the posterior. Since we assume all \( q(T^q_t) \) to share the same covariance structure, mode finding on this distribution equates to applying meanshift [10] with a Gaussian kernel of covariance \( \Sigma \). Note that since \( T^q_t \) 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 \( \hat{T}^q_t \) are first converted to a 6-vector \( p = \log(\hat{T}^q_t) \in \mathfrak{se}(3) \) in Lie algebra through logarithm mapping, and meanshift is applied to the 6-vector space.

**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.

**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 \( X = \{ X_t \mid t = 1 \cdots t_N \} \), and our output will be the camera poses of \( T = \{ T_t \mid t = 1 \cdots t_N \} \) as well as the depth map \( \theta \) of the first frame. Usually, 4-8 optical flows per batch are used. Firstly, VOLDOR initializes all \( W \) to one and \( T_t \) is initialized from epipolar geometry estimated from \( X_t \) using a least median of squares estimator [68] or, alternatively, from previous estimates if available (i.e. overlapping consecutive frame batches). Then, \( \theta \) is obtained from two-view triangulation using \( T_t \) and \( X_t \). 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 \( \lambda \) in Eq. (8) to 0.15, \( \gamma \) in Eq. (11) to 0.9. The Gaussian kernel covariance matrix \( \Sigma \) 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]
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.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>VISO2-M</th>
<th>MLM-SFM</th>
<th>VOLDOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg.</td>
<td>10.85</td>
<td>0.0249</td>
<td>2.03</td>
</tr>
<tr>
<td>Trans. (%)</td>
<td>0.0045</td>
<td>0.0042</td>
<td>1.32</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Trans. (%)</th>
<th>Rot. (deg/m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VISO2-M [26]</td>
<td>11.94</td>
<td>0.0234</td>
</tr>
<tr>
<td>MLM-SFM [74] [75]</td>
<td>2.54</td>
<td>0.0057</td>
</tr>
<tr>
<td>PVS/M2 [19] [17] [18]</td>
<td>2.05</td>
<td>0.0051</td>
</tr>
<tr>
<td>VBO [62]</td>
<td>1.76</td>
<td>0.0036</td>
</tr>
<tr>
<td>VOLDOR (Ours)</td>
<td>1.65</td>
<td>0.0050</td>
</tr>
</tbody>
</table>

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 $W^j = \sum_i W^j_i$.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>ORB-SLAM2 (RGB-D)</th>
<th>DVO-SLAM (RGB-D)</th>
<th>DSO (Mono)</th>
<th>GA-Net-15 (Mono)</th>
<th>Ours (Mono)</th>
</tr>
</thead>
<tbody>
<tr>
<td>fr1/desk</td>
<td>0.0163</td>
<td>0.0185</td>
<td>0.0168</td>
<td>0.0133</td>
<td></td>
</tr>
<tr>
<td>fr1/desk2</td>
<td>0.0162</td>
<td>0.0238</td>
<td>0.0188</td>
<td>0.0150</td>
<td></td>
</tr>
<tr>
<td>fr1/room</td>
<td>0.0102</td>
<td>0.0117</td>
<td>0.0108</td>
<td>0.0090</td>
<td></td>
</tr>
<tr>
<td>fr2/desk</td>
<td>0.0045</td>
<td>0.0068</td>
<td>0.0048</td>
<td>0.0053</td>
<td></td>
</tr>
<tr>
<td>fr2/xyz</td>
<td>0.0034</td>
<td>0.0055</td>
<td>0.0025</td>
<td>0.0034</td>
<td></td>
</tr>
<tr>
<td>fr3/office</td>
<td>0.0046</td>
<td>0.0102</td>
<td>0.0050</td>
<td>0.0045</td>
<td></td>
</tr>
<tr>
<td>fr3/inst</td>
<td>0.0079</td>
<td>0.0073</td>
<td>0.0087</td>
<td>0.0071</td>
<td></td>
</tr>
</tbody>
</table>

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

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, different rigidness probabilities, where $W^j = \sum_i W^j_i$.

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.
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 \( W^f \)) are used. Performance across different residual models tends to converge as the number of reliable samples increases (high \( W^f \)), 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 \( W^f \)) 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


73] Shiyu Song and Mamohand 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


