INFORMATION-THEORETIC ODOMETRY LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we propose a unified information-theoretic framework for odometry learning, a crucial component of many robotics and vision tasks such as navigation and virtual reality where 6-DOF poses are required in real time. We formulate this problem as optimizing a variational information bottleneck objective function, which eliminates pose-irrelevant information from the latent representation. The proposed framework provides an elegant tool for performance evaluation and understanding in information-theoretic language. Specifically, we bound the generalization errors of the deep information bottleneck framework and the predictability of the latent representation. These provide not only a performance guarantee but also practical guidance for model design, sample collection, and sensor selection. Furthermore, the stochastic latent representation provides a natural uncertainty measure without the needs for extra structures or computations. Experiments on two well-known odometry datasets demonstrate the effectiveness of our method.

1 INTRODUCTION

Odometry aims to predict six degrees of freedom (6-DOF) poses from motion sensors. It is a fundamental component of a wide variety of robotics and vision tasks including simultaneous localization and mapping (SLAM), automatic navigation, and virtual reality (Durrant-Whyte & Bailey, 2006; Fuentes-Pacheco et al., 2015; Taketomi et al., 2017). In particular, visual and visual-inertial odometry have attracted a lot of attention over recent years due to the low cost and easy setup of cameras and inertial measurement unit (IMU) sensors. Odometry is challenging due to the difficulties to model the complexity and diversity of real-world scenarios from a limited number of on-board sensors. Furthermore, since odometry is essentially a time-series prediction problem, how to properly handle time dependency and environment dynamics presents further challenges.

Current visual odometry solutions fall into geometry constraint based methods and deep learning based methods. Though successful in many real-world scenarios, the performance of geometry-based approaches can be limited when their underlying assumptions, such as static environments, discriminative visual features and brightness constancy, are violated. Furthermore, their hard-coded systems require extensive parameter tuning under different environments and non-trivial synchronization and initialization. Deep learning-based methods have recently attracted interest by learning from large-scale datasets (Wang et al., 2017; Clark et al., 2017; Xue et al., 2019; Yang et al., 2020). Well-trained deep networks effectively capture the inherent complexity and diversity of the training data, thus holding promise for addressing the limitations of geometry based approaches.

Although existing deep odometry learning methods have performed competitively against their geometry counterparts, they still fail to satisfy some basic requirements. First, due to the broad range of scenarios where odometry is required, odometry systems are expected to be easily compatible with various configurations and settings such as multiple sensors and dynamic environments. Besides, the common existence of data degeneration, such as from hardware malfunctions and unexpected occlusions, requires a safe and robust system in which a proper uncertainty measure is desirable for self-awareness of anomalies and system bias. Moreover, theoretical analyses of current black box deep odometry models, such as generalizability on unseen test data and extendibility to extra sensors, are still obscure but essential for understanding and assessing the model performance.

Here we devise a unified odometry learning framework from an information-theoretic perspective, which well addresses the above issues. Our work is motivated by the recent successes of deep variational inference and learning theory based on mutual information (MI). Specifically, we translate the odometry problem to optimizing an information bottleneck (IB) objective function where the latent representation is formulated as a bottleneck between observations and poses. In doing so, we eliminate the pose-irrelevant information from the latent representation to achieve better generalizability. Modeling by MI constraints provides a flexible way to account for different aspects of the problem and quantify their effectiveness in information-theoretic language. This framework is also attractive in that the operations are performed on the probabilistic distribution of the latent representation, which naturally provides an uncertainty measure for interrogating data quality and system bias.

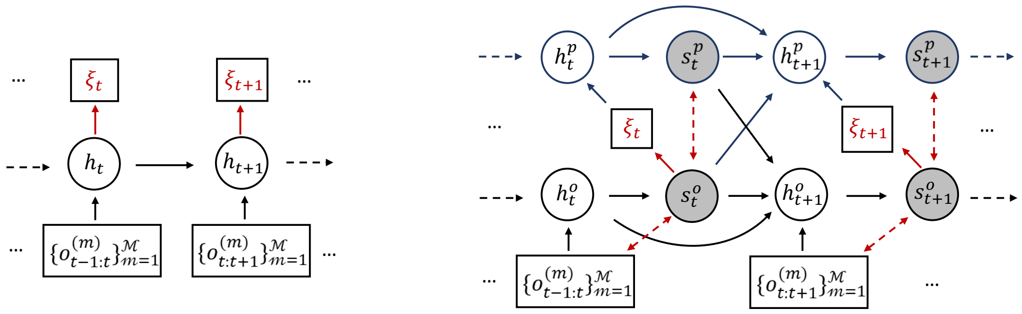
More importantly, the information-theoretic formulation allows us to leverage information theory to investigate the theoretical properties of the proposed method. Our theoretical findings not only benefit the evaluation of the model performance but also provide insights for subsequent research. We obtain a theoretical guarantee of the proposed framework by deriving an upper bound of the generalization error w.r.t. the IB objective function under mild network and loss function conditions. We show that the latent space dimensionality also bounds the generalization error, providing a theoretical explanation for the complexity-overfitting trade-off in the latent representation space. When test data is biased, our result shows that the growing rate of d should not exceed that of $n/log(n)$, where d is the latent space dimensionality and n is the sample size. We further quantify the usefulness of a latent representation for pose prediction using the MI between the representation and poses. In doing so, we prove a lower bound for this MI given extra sensors, which reveals the conditions required for a sensor to theoretically guarantee a performance gain. It is noteworthy that our theoretical results hold not only for the odometry problem but also for a wider variety of problems that share the same Markov chain assumption and the IB objective function. A connection between our information-theoretic framework and geometry based methods is further established for deeper insights.

The main contributions of this paper are (1) we propose information-theoretic odometry learning by leveraging the IB objective function to eliminate pose-irrelevant information from the latent representation; (2) we develop the theoretical performance guarantee of the proposed framework by deriving upper bounds on the generalization error w.r.t. IB and the latent space dimensionality as well as a lower bound on the MI between the latent representation and poses; and (3) we empirically verify the effectiveness of our method on the well-known KITTI and EuRoC datasets and show how the intrinsic uncertainty benefits failure detection and inference refinement.

2 Related Work

Deep representation for odometry learning: Leveraging deep neural networks to learn compact feature representation from high-dimension sensor data has been proven effective for odometry. Kendall et al. (2015) proposed PoseNet by using neural networks for camera relocalization, based upon which Wang et al. (2017) introduced a recurrent module to model the temporal correlation of features for visual odometry. Subsequently, Xue et al. (2019) further considered a memory and refinement module to address the prediction drift caused by error accumulation. Recently, deep learning-based odometry has also been extended to the multi-sensor configuration. Clark et al. (2017) extended the DeepVO framework to incorporate IMU data by leveraging an extra recurrent network for learning better feature representation. A recent study by Chen et al. (2019) investigated more effective and robust sensor fusion via soft and hard attention for visual-inertial odometry. Apart from end-to-end learning, there are also trends in unsupervised learning (Zhou et al., 2017; Yin & Shi, 2018; Ranjan et al., 2019; Bian et al., 2019) and the combination of learned features with geometry methods (Zhan et al., 2019; Yang et al., 2020). We refer readers to Chen et al. (2020) for a more detailed discussion of current methods. These deep odometry learning methods have achieved promising performance. However, theoretical understandings remain obscure: (1) how to learn a compact representation with a theoretically guaranteed generalizability when test data is biased and (2) in what conditions extra sensors can benefit the pose prediction problem.

Information bottleneck: Information bottleneck (IB) provides an appealing tool for deep learning by learning an informative and compact latent representation (Tishby et al., 2000; Tishby & Zaslavsky, 2015; Shwartz-Ziv & Tishby, 2017). To address the intractability of MI calculation, Alemi et al. (2017) proposed to optimize a variational bound of IB for deep learning, which was successfully applied to many tasks including dynamics learning (Hafner et al., 2020), task transfer (Goyal et al., 2019), and network compression (Dai et al., 2018). Partly inspired by these developments, we for the first time propose an IB-based framework for odometry learning and derive an optimizable



(a) Classic Odometry Learning

(b) Deterministic-Stochastic IB Odometry Learning

Figure 1: (a) The classic learning-based odometry framework, where 6-DOF poses are directly predicted from deterministic latent representations. (b) The proposed information bottleneck (IB) framework for odometry learning. h and s are the deterministic and stochastic components, respectively. Superscripts o and p represent the observation- and pose-level transition models. Red solid arrows denote the pose regressor and red dashed arrows denote the bottleneck constraints. Output arrows from a shaded stochastic representation represent samples from the learned latent distribution.

variational bound for this sequential prediction problem. The derivation can be more delicate if we incorporate more constraints, potentially from geometric and kinematic insights. We further adopt the determinisitic-stochastic separation as in Chung et al. (2015); Hafner et al. (2019; 2020), while ours differs in that our derivation of the variational bound allows modeling two transition models separately, each with a determinisic component to improve model capacity. Moreover, though IB-based methods have shown to be effective for learning a compact representation, the underpinning generalizability theory remains unclear. The generalization error bounds for general learning algorithms have been studied in Xu & Raginsky (2017) in information-theoretic language. This work was subsequently extended by Zhang et al. (2018) to explain the generalizability of deep neural networks. However, their results are not applicable to the IB-based methods, which will be addressed in this paper.

Uncertainty modeling for odometry learning: Modeling uncertainty to deal with extreme cases like hardware malfunctions and unexpected occlusions, is crucial for a reliable and robust odometry system. It can be categorized into model-intrinsic epistemic uncertainty and data-dependent aleatoric uncertainty, which have been studied in the Bayesian deep learning literature (MacKay, 1992; Gal & Ghahramani, 2016; Kendall & Gal, 2017). For odometry, Wang et al. (2018) and Yang et al. (2020) captured the aleatoric uncertainty by imposing a probabilistic distribution on poses and used the second moment prediction as an uncertainty measure. Recently, Loquercio et al. (2020) showed that a combined epistemic-aleatoric uncertainty framework (Kendall & Gal, 2017) could improve the performance on several robotics tasks such as motion and steering angle predictions. In contrast to them, our framework provides a built-in and efficient uncertainty measure that accounts for both uncertainty types. We empirically demonstrate how to use this uncertainty measure to evaluate data quality and system biases. Accordingly, we propose a refined inference procedure that discards highly uncertain results to improve pose prediction accuracy.

3 Information-Theoretic Odometry Learning

Odometry aims to predict the relative 6-DOF pose ξ_t between two consecutive observations $\{o^{(m)}_{t-1:t}\}^{\mathcal{M}}_{m=1}$ from $\mathcal{M}$ sensors (e.g. camera, IMU and lidar), where t is the time index. This pose prediction problem can be formulated as $\xi_t = g(\{o^{(m)}_{t-1:t}\}^{\mathcal{M}}_{m=1}, \Theta)$, where g is the mapping function of an odometry system and Θ is the parameter set of g. Classic deep odometry learning methods model g by neural networks and learn Θ from training data. Furthermore, they usually use a recurrent module to model the motion dynamics of the observation sequence. Figure 1(a) shows a typical procedure shared by representative deep odometry learning methods.

In many settings, observations are of high dimensionality, such as images and lidar 3D points. Geometry methods use low-dimensional features to represent observations, while learning-based methods learn a representation from training data. However, both features may contain pose-irrelevant information that are specific to certain sensor domain. Retaining such information encourages the model to overfit the training data and yield poor generalization performance. Since parsimony is preferred in machine learning, it is expected to eliminate those pose-irrelevant information.

To this end, we tackle this problem by explicitly introducing a constraint on the pose-irrelevant information. Specifically, we quantify the pose-irrelevance and the usefulness of a latent representation for pose prediction from an information-theoretic perspective. By assuming the latent representation s_t at time t is drawn from a Gaussian distribution, the MI $I(\{o^{(m)}_{1:T}\}^{\mathcal{M}}_{m=1}||s_{1:T}|\xi_{1:T})$ and the MI $I(\xi_{1:T}||s_{1:T})$ can provide quantitative measures for the aforementioned two aspects. Accordingly, given a sequence of observations $\{o^{(m)}_{1:T}\}^{\mathcal{M}}_{m=1}$ and pose annotations $\xi_{1:T}$ from time 1 to T, our information-theoretic odometry learning problem is:

$$max_{\Theta}\ \mathcal{J}(\Theta) = I(\xi_{1:T}||s_{1:T}) - \gamma I(\{o^{(m)}_{1:T}\}^{\mathcal{M}}_{m=1}||s_{1:T}|\xi_{1:T}), \tag{1}$$

where γ controls the trade-off between the two MI terms. By Equation 1, the latent representation $s_{1:T}$ essentially provides an information bottleneck between poses and observations, which eliminates pose-irrelevant information from the observations. Due to the high dimensionality of the observation space, it is non-trivial to calculate the two MI. Thus we optimize a variational lower bound instead:

$$\mathcal{J}(\Theta) \geq \mathcal{J}'(\Theta) = E_{s_{1:T},\{o^{(m)}_{1:T}\}^{\mathcal{M}}_{m=1},\xi_{1:T}}\left[\sum_{t=1}^{T}(\mathcal{J}^{pose}_t - \gamma\mathcal{J}^{bottleneck}_t)\right], \tag{2}$$

$$\mathcal{J}^{pose}_t = log\ q_{\theta}(\xi_t|s_t), \tag{3}$$

$$\mathcal{J}^{bottleneck}_t = D_{KL}[p_{\phi}(s_t|\{o^{(m)}_{t-1:t}\}^{\mathcal{M}}_{m=1}, s_{t-1})||q_{\varphi}(s_t|\xi_t, s_{t-1})]. \tag{4}$$

The detailed derivation is provided in the Appendix. This lower bound consists of a variational pose regressor $q_{\theta}(\xi_t|s_t)$, an observation-level transition model $p_{\phi}(s_t|\{o^{(m)}_{t-1:t}\}^{\mathcal{M}}_{m=1}, s_{t-1})$, and a pose-level transition model $q_{\varphi}(s_t|\xi_t, s_{t-1})$, all modeled by neural networks. For simplicity, we denote the representations from the observation-level and pose-level transition models s^o_t and s^p_t, respectively. In practice, s^o_t is used for the pose regressor. Intuitively, minimizing the KL divergence in Equation 4 forces the distribution of s^o_t to approximate that of s^p_t which does not encode the observation information at time t, thus regularizing s^o_t for containing pose-irrelevant information.

Stochastic-only transition models, however, may compromise model performance due to uncertainty accumulation during the sampling process. To address this problem, we further introduce a deterministic component according to Chung et al. (2015) and Hafner et al. (2019). In doing so, we reformulate the two transition models in the KL divergence in Equation 4 as:

$$\text{observation-level} : \quad p_{\phi}(s^o_t|h^o_t),\ h^o_t = f^o(h^o_{t-1}, \{o^{(m)}_{t-1:t}\}^{\mathcal{M}}_{m=1}, s^o_{t-1}, s^p_{t-1}), \tag{5}$$

$$\text{pose-level} : \quad q_{\varphi}(s^p_t|h^p_t),\ h^p_t = f^p(h^p_{t-1}, \xi_t, s^o_{t-1}, s^p_{t-1}). \tag{6}$$

We use two deterministic functions f^o and f^p for observation- and pose-level transtions, respectively, which are modeled by recurrent neural networks. Both s^o_{t-1} and s^p_{t-1} are used for the two deteministic transition functions to help to reduce the KL divergence between the distributions of s^o_t and s^p_t. Ground-truth 6-DOF poses are fed into f^p during the training phase while for testing we use predicted poses to provide a runtime estimate of s^p_t. Figure 1(b) shows the overall framework of our method.

Since we model the latent representation in the probabilistic space, the variance of the latent representation naturally provides an uncertainty measure. We empirically show how this intrinsic uncertainty reveals data quality and system bias in Section 5.3. Of note is that it is straight forward to extend the proposed information-theoretic framework to different problem settings. We can add arbitrary linear MI constraints into the proposed objective and derive similar variational bounds to satisfy different requirements such as dynamics-awareness in complex environments.

4 Theoretical Analysis

Formulating a problem in information-theoretic language enables us to analyze the proposed method by exploring elegant tools in information theory (Cover & Thomas, 1991) and related results in learning theory (Xu & Raginsky, 2017; Zhang et al., 2018). In this work, we show that the MI between the bottleneck and observations as well as the latent space dimensionality upper bound the expected generalization error, which provides not only insights into the generalizability of the method but also a performance guarantee. To our knowledge, this is the first time that such generalization bounds have been derived for IB by using a general loss function other than cross-entropy (Vera et al., 2018). By replacing the general loss function with the cross-entropy, our bound is tighter than that obtained by Vera et al. (2018) in terms of the sample size. We further derive a lower bound on the MI between the latent representation and poses given extra sensors, which suggests what features make a sensor useful for pose prediction in information-theoretic language. The connection between information bottleneck and geometry based methods is also established to provide further insights.

4.1 Generalization Bound for Information Bottleneck

Xu & Raginsky (2017) and Zhang et al. (2018) obtained the generalization bound w.r.t. the MI between input data X and learning parameters Θ for general learning algorithms and neural networks. However, what IB regularizes is the MI between X and the latent representation. To derive a generalization bound for the IB objective function, we first prove a relationship between these two kinds of MI in Lemma 2 under the Markov chain $X \rightarrow S \rightarrow \xi$, an underlying assumption for IB.

Lemma 1. *If $X \rightarrow S \rightarrow \xi$ forms a Markov chain and assume $\xi = g(X, \Theta)$ is a one-to-one function w.r.t. X and Θ, then we have*

$$I(X, S) \geq I(X, \xi) = I(X, \Theta) + E_\theta[H(X|\theta)] \geq I(X, \Theta). \tag{7}$$

Lemma 2 enables us to extend the generalizability results for neural networks regarding $I(X, \Theta)$ (Zhang et al., 2018) to the IB setting, leading to the following theoretical counterpart:

Theorem 1. *Assuming $X \rightarrow S \rightarrow \xi$ is a Markov chain, the loss function $l(X, \Theta)$ is sub-σ-Gaussian distributed[1] and the prediction function $\xi = g(X, \Theta)$ is a one-to-one function w.r.t. the input data and network parameters Θ, we have the following upper bound for the expected generalization error:*

$$E[R(\Theta) - R_T(\Theta)] \leq exp(-\frac{L}{2}log\frac{1}{\eta})\sqrt{\frac{2\sigma^2}{n}I(X, S)}, \tag{8}$$

where L, η, and n are the effective number of layers causing information loss, a constant smaller than 1, and the sample size, respectively. $R(\Theta) = E_{X\sim D}[l(X, \Theta)]$ is the expected loss value given Θ and $R_T(\Theta) = \frac{1}{n}\sum_{i=1}^{n} l(X_i, \Theta)$ is a sample estimate of $R(\Theta)$ from the training data.

The difference between our result and previous work is that we bound the generalization error by $I(X, S)$ which is minimized in Equation 1 rather than $I(X, \Theta)$ which is hard to evaluate. By Theorem 3, we show that minimizing the MI between the bottleneck and observations tightens the upper bound on the expected generalization error and thus provides a theoretical performance guarantee. It is worth noting that our theoretical results apply not only to our odometry learning setting but also to a wider variety of tasks which use the IB method. This bound also implies that a larger sample size and a deeper network lead to better generalization performance, which is consistent with the results in Xu & Raginsky (2017) and Zhang et al. (2018). The detailed proof and further discussion of Lemma 2 and Theorem 3 can be found in the Appendix.

4.2 Generalization Bound for Latent Dimensionality

We further investigate the generalizability w.r.t. model complexity in terms of the cardinality and dimensionality of the latent representation space under the IB framework.

Corollary 1. *Given the same assumptions in Theorem 3 and let |S| be the cardinality of the latent representation space, we have*

$$E[R(\Theta) - R_T(\Theta)] \leq exp(-\frac{L}{2}log\frac{1}{\eta})\sqrt{\frac{2\sigma^2}{n}log|S|}. \tag{9}$$

[1] Recall that a random variable l is sub-σ-Gaussian distributed if $E[e^{\lambda(l-E[l])}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$, $\forall\, \lambda \in R$.

It is well recognized that a large model complexity can impair the model generalizability. We reveal this complexity-overfitting trade-off in Corollary 3, where the expected generalization error is upper bounded by the cardinality of the latent representation space. Considering the model design and sample collection, Corollary 3 indicates that the growing rate of $log|S|$ should not exceed that of n to avoid an exploded generalization error bound.

Corollary 2. *Given the same assumptions in Theorem 3 and assume S lies in a d-dimensional subspace of the latent representation space, $sup_{s_i \in S_i} ||s_i|| \leq M, \forall i \in [1, d]$ and S can be approximated by a densely quantized space, the following generalization bound holds:*

$$E[R(\Theta) - R_T(\Theta)] \leq exp(-\frac{L}{2}log\frac{1}{\eta})\sigma\sqrt{\frac{dlog(d)}{n} + 2log(2M)\frac{d}{n} + \frac{d}{n/log(n)}}. \qquad (10)$$

In practice, it is usually difficult to evaluate $log|S|$ in Corollary 3 numerically. Therefore, we leverage the quantization trick used in Xu & Raginsky (2017) to reduce the upper bound to a function w.r.t. the dimensionality d of the latent representation space. The result is given in Corollary 4, which suggests that the growing rate of d should not exceed that of $n/log(n)$. It is worth noting that this result holds not only for IB but also for a broader range of encoder-decoder models under the Markov chain assumption on $X \to S \to \xi$.

4.3 Predictability Bound for Extra Sensors

Odometry performance is highly dependent on the sensors deployed, yet it remains non-trivial to select informative sensors that guarantee a performance gain. In this section, we address this problem using information-theoretic language under our proposed framework.

Theorem 2. *If* $(\{o^{(m)}\}_{m=1}^{\mathcal{M}},\ o^{(\mathcal{M}+1)}) \to S \to \xi$ *forms a Markov chain, then we have,*

$$I(\xi||S) \geq I(\xi||\{o^{(m)}\}_{m=1}^{(\mathcal{M})}) + I(\xi||o^{(\mathcal{M}+1)}|\{o^{(m)}\}_{m=1}^{\mathcal{M}}) - I(o^{(\mathcal{M}+1)}||\{o^{(m)}\}_{m=1}^{\mathcal{M}}|\xi). \qquad (11)$$

Theorem 4 suggests that if a new sensor $o^{(\mathcal{M}+1)}$ is useful for pose prediction, the MI between $o^{(\mathcal{M}+1)}$ and poses given existing sensors should be large. Meanwhile, it is preferred to have a small MI between $\{o^{(m)}\}_{m=1}^{(\mathcal{M})}$ and $o^{(\mathcal{M}+1)}$ given pose information. In other words, a heterogenous sensor that shares little pose-irrelevant information with existing sensors is desirable. The information gain between $I(\xi||o^{(\mathcal{M}+1)}|\{o^{(m)}\}_{m=1}^{\mathcal{M}})$ and $I(o^{(\mathcal{M}+1)}||\{o^{(m)}\}_{m=1}^{\mathcal{M}}|\xi)$ provides a theoretical guarantee for the performance of the learned latent representation.

4.4 Connection with Geometry Methods

More generally, an odometry system can be modeled as $h(z_{k,j}, v_k, \check{x_k}) \to (\hat{x_k}, p_j)$ where $z_{k,j}, v_k, \check{x_k}, \hat{x_k}$ and p_j are observations, noise, prior pose, posterior pose, and latent state, respectively. At this level, the bottleneck MI $I(z_{k,j}, v_k||p_j|\hat{x_k}) = H[h(z_{k,j}, v_k, \check{x_k})|\hat{x_k}] - H[h(z_{k,j}, v_k, \check{x_k})|\hat{x_k}, z_{k,j}, v_k]$ is the extra entropy (ΔH) introduced by $(z_{k,j}, v_k)$, which differs for different h. Factor graph based methods use optimization over L2 costs as h, where p_j is inferred landmark and a Gaussian noise is assumed. ΔH in this case is implied in the noise variance which corresponds to the pre-specified weight of each cost function. Learning based methods learn h from data where p_j is the latent feature. Minimizing ΔH means reducing the uncertainty from noise and inexact learned function forms. The same analysis applies to kinematic function for $\check{x_k}$. Besides, filter-based methods can also be included in by following the same logic. Take the kinematics part of Kalman filter (linear Gaussian system) as an example: $\check{x_k} = A_k \hat{x_{k-1}} + u_k + w_k$, where the prior $\check{x_k}$ is the latent state and the variance of $\hat{x_{k-1}}$ and w_k are $\hat{\Sigma_{k-1}}$ and R, respectively. Then $I(u_k, w_k||\check{x_k}) = \frac{1}{2}ln(|A_k \hat{\Sigma_{k-1}} A_k^T + R|/|A_k \hat{\Sigma_{k-1}} A_k^T|)$, suggesting that a smaller bottleneck MI corresponds to a relatively smaller noise variance.

5 Experiments

We tested our method on the well-known KITTI (Geiger et al., 2013) and EuRoC (Burri et al., 2016) datasets. Since most existing supervised methods are not open source, we re-implemented

Table 1: Test results on KITTI and EuRoC. We report the average RMSEs for translation and rotation, respectively. †: Results of MSCKF on KITTI and OKVIS on EuRoC are from Chen et al. (2019).

Model	KITTI		EuRoC	
	$t(m)$	$r(^o)$	$t(m)$	$r(^o)$
DeepVO	0.0658	0.0942	0.0323	0.2114
InfoVO	**0.0607**	**0.0869**	**0.0310**	**0.2061**
MSCKF/OKVIS†	0.116	0.044	0.0283	**0.0402**
VINet	0.0629	0.0453	0.0281	0.0729
SoftFusion	0.0629	0.0517	0.0281	0.0672
HardFusion	0.0618	0.0447	0.0285	0.0740
InfoVIO	0.0580	**0.0416**	0.0276	0.0744
SoftInfoVIO	0.0618	0.0438	**0.0272**	0.0743
HardInfoVIO	**0.0559**	0.0454	0.0291	0.0763

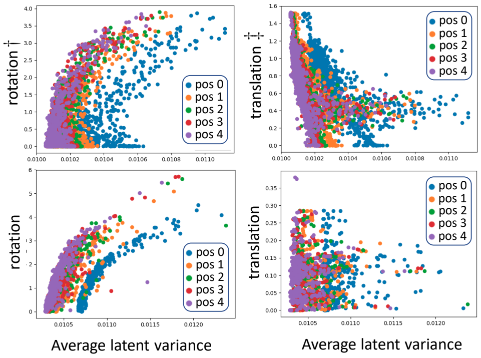


Figure 2: Visualization of intrinsic uncertainty vs. rotation and translation. Top: KITTI. Bottom: EuRoC. †: rotation for turning left or right. ‡: translation towards the forward direction. pos: the evaluated position.

the representative state-of-the-art methods including DeepVO (Wang et al., 2017), VINet (Clark et al., 2017), and two attention-based visual-inertial methods recently proposed by Chen et al. (2019), namely, SoftFusion and HardFusion, as our baselines. All models shared the same network architecture for a fair comparison. We also conducted extensive ablation studies on the deterministic component, the sample size, extra sensors and the intrinsic uncertainty measure.

5.1 Datasets and Experimental Settings

The KITTI odometry dataset consists of 11 real-world car driving videos and calibrated ground-truth 6-DOF pose annotations. The EuRoC dataset was instead collected from a MAV in two buildings, resulting in 11 sequences of different difficulties by manually adjusted obstacles. For visual-inertial experiments, we manually aligned the 100 Hz IMU records in the raw KITTI dataset to the 10 Hz image sequences using the corresponding timestamps. The image and IMU sequences in EuRoC were downsampled to 10 Hz and 100 Hz, respectively. We split the training and test datasets following the recent work by Chen et al. (2019). Our implementation was based on PyTorch (Steiner et al., 2019) and we will release the source code package and the trained models. We used GRU (Cho et al., 2014) to model the deterministic transitions and IMU records. Pretrained FlowNet was used to extract features from image data (Dosovitskiy et al., 2015; Ilg et al., 2017). The other parts were modeled by MLP layers. More details are given in the Appendix.

5.2 Main Results

We implemented our visual-inertial framework using three fusion strategies proposed in Chen et al. (2019), namely InfoVIO, SoftInfoVIO, and HardInfoVIO. We also included two traditional visual-inertial odometry methods for comparison, i.e., OKVIS (Leutenegger et al., 2015) for EuRoC and MSCKF (Hu & Chen, 2014) for KITTI. OKVIS is not used for KITTI due to the lack of accurate time synchronization between images and IMU data. Following Sturm et al. (2012) and Chen et al. (2019), we report the average root mean squared errors (RMSEs) of translation and rotation. The results are given in Table 1. Our results support the effectiveness of IB w.r.t. the generalizability to test data. Specifically, our basic models (InfoVO/InfoVIO) outperformed all baselines w.r.t. both metrics on KITTI and the translation error on EuRoC. Visual odometry models performed well for translation prediction while incorporating IMU significantly improved the rotation results. Since the MAV trajectories are challenging w.r.t. rotation, the traditional method (OKVIS) still outperformed the other methods, although our result was competitive with the other learning-based baselines. Our re-implementation achieved a better result on KITTI compared with Chen et al. (2019) but the performance on EuRoC degraded. Since EuRoC is much more challenging than KITTI, reducing the performance gap on EuRoC may require more carefully designed training strategies. Comparisons between the two datasets are given in the Appendix. We will fine-tune on EuRoC in a future study.

Ablation studies: Extensive ablation studies were conducted to examine the effects of (1) the deterministic component, (2) sample size and (3) extra sensors. Key observations include: (1)

Table 2: Results on KITTI by evaluating at different positions in a clip.

$t(m)$	pos-0	pos-1	pos-2	pos-3	pos-4
DeepVO	0.0734	0.0681	0.0661	**0.0658**	0.0659
InfoVO	0.0689	0.0631	0.0618	0.0608	**0.0604**
VINet	0.0683	0.0645	0.0645	0.0632	**0.0615**
InfoVIO	0.0671	0.0602	0.0586	0.0580	**0.0572**
$r(^o)$	pos-0	pos-1	pos-2	pos-3	pos-4
DeepVO	0.0970	0.0949	**0.0939**	0.0940	0.0951
InfoVO	0.0904	0.0881	0.0871	**0.0869**	0.0872
VINet	0.0463	0.0455	**0.0454**	**0.0454**	0.0456
InfoVIO	0.0427	**0.0417**	0.0420	0.0420	0.0421

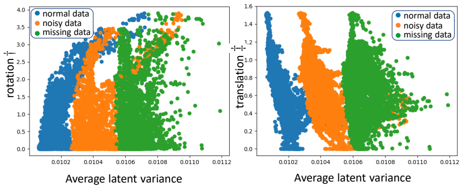


Figure 3: Visualization of intrinsic uncertainty for degraded data on KITTI. †: rotation for turning left or right. ‡: translation towards the forward direction.

Without the deterministic component, both translation and rotation performance dropped significantly; (2) A larger sample size reduces both the uncertainty and prediction errors; and (3) IMU is more ‘useful’ than cameras for rotation prediction while cameras are more crucial than IMU for translation prediction, according to the discussions on Theorem 2. Detailed results are provided in the Appendix.

5.3 What Does the Intrinsic Uncertainty Mean?

We next used the average variance of the stochastic latent representation as an intrinsic uncertainty measure and empirically showed how this uncertainty reveals the system properties and data degradation. We found some interesting relationships between the uncertainty and poses, i.e., larger turning angles and smaller forward distances lead to higher uncertainty, as shown in Figure 2. The reason can be that a large forward parallax provides more distinctive matching features for pose prediction while a large turning angle instead dramatically reduces the shared visible areas and results in difficulties to achieve accurate prediction. Our analysis suggests a practical data collection guideline, i.e., augmenting the uncertain parts of the pose distribution.

Uncertainty w.r.t. the evaluated position in a clip: We trained and evaluated the odometry model in a clip-wise manner. Surprisingly, the evaluated position for a frame-pair in consecutive clips significantly affected the intrinsic uncertainty, as shown in Figure 2. This makes sense in that when evaluated at a latter position of a clip, the prediction model can leverage more information accumulated from former observations, thus leading to more confident predictions. In Table 2, we show that, in general, a larger uncertainty results in a higher prediction error. The result also holds for the deterministic DeepVO and VINet baselines, implying that this is a structural system problem in the clip-wise recurrent models. Based on this observation, we propose a simple refinement strategy that eliminates results from the most uncertain position (pos-0) and averagely ensembles the rest. We report the refined evaluation results for all models in our main results and ablation studies.

Failure-awareness: We show that the intrinsic uncertainty measure is failure-aware, which is crucial for a robust odometry system. We present the results from noisy and missing data on KITTI in Figure 3. Our model becomes more uncertain as the data degrades. The uncertainty reaches highest when the data is missing, as expected. More interestingly, for InfoVIO the quality of IMU data dominates the uncertainty, implying that current learned models focus more on IMU data and that a better image encoder is desirable. Degradation details and extended results are given in the Appendix.

6 Conclusion

This paper targets odometry learning by proposing an information-theoretic framework that leverages an IB-based objective function to eliminate the pose-irrelevant information. A recurrent deterministic-stochastic transition model is introduced to facilitate the modeling of time dependency of the observation sequences. The proposed framework can be easily extended to different problem settings and provide not only an intrinsic uncertainty measure but also an elegant theoretical analysis tool for evaluating the system performance. We derive generalization error bounds for the IB-based method and a predictability lower bound for the latent representation given extra sensors. They provide theoretical performance guarantees for the proposed framework, and more generally, information-bottleneck based methods. Extensive experiments on KITTI and EuRoC support our discoveries.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR 2017 : International Conference on Learning Representations 2017*, 2017.

JiaWang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *NeurIPS 2019 : Thirty-third Conference on Neural Information Processing Systems*, pp. 35–45, 2019.

Lars Buesing, Theophane Weber, Sébastien Racanière, S. M. Ali Eslami, Danilo Jimenez Rezende, David P. Reichert, Fabio Viola, Frederic Besse, Karol Gregor, Demis Hassabis, and Daan Wierstra. Learning and querying fast generative models for reinforcement learning. *arXiv preprint arXiv:1802.03006*, 2018.

Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35(10):1157–1163, 2016.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *international conference on learning representations*, 2019(12):124018, 2017.

Changhao Chen, Stefano Rosa, Yishu Miao, Chris Xiaoxuan Lu, Wei Wu, Andrew Markham, and Niki Trigoni. Selective sensor fusion for neural visual-inertial odometry. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10542–10551, 2019.

Changhao Chen, Bing Wang, Chris Xiaoxuan Lu, Niki Trigoni, and Andrew Markham. A survey on deep learning for localization and mapping: Towards the age of spatial machine intelligence. *arXiv preprint arXiv:2006.12567*, 2020.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, pp. 2980–2988, 2015.

Ronald Clark, Sen Wang, Hongkai Wen, Andrew Markham, and Niki Trigoni. Vinet: Visual inertial odometry as a sequence to sequence learning problem. In *Thirty-First AAAI Conference on Artificial Intelligence*, pp. 3995–4001, 2017.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. 1991.

Bin Dai, Chen Zhu, and David Wipf. Compressing neural networks using the variational information bottelneck. In *ICML 2018: Thirty-fifth International Conference on Machine Learning*, pp. 1135–1144, 2018.

Alexey Dosovitskiy, Philipp Fischery, 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 *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2758–2766, 2015.

H. Durrant-Whyte and T. Bailey. Simultaneous localization and mapping: part i. *IEEE Robotics & Automation Magazine*, 13(2):99–110, 2006.

Jorge Fuentes-Pacheco, José Ruiz-Ascencio, and Juan Manuel Rendón-Mancha. Visual simultaneous localization and mapping: a survey. *Artificial Intelligence Review*, 43(1):55–81, 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In *ICML'16 Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, pp. 1050–1059, 2016.

A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

Anirudh Goyal Alias Parth Goyal, Riashat Islam, DJ Strouse, Zafarali Ahmed, Hugo Larochelle, Matthew Botvinick, Sergey Levine, and Yoshua Bengio. Infobot: Transfer and exploration via the information bottleneck. In *ICLR 2019 : 7th International Conference on Learning Representations*, 2019.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *ICML 2019 : Thirty-sixth International Conference on Machine Learning*, pp. 2555–2565, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR 2020 : Eighth International Conference on Learning Representations*, 2020.

Jwu-Sheng Hu and Ming-Yuan Chen. A sliding-window visual-imu odometer based on tri-focal tensor geometry. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3963–3968, 2014.

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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1647–1655, 2017.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision. In *NIPS'17 Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5580–5590, 2017.

Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2938–2946, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR 2014 : International Conference on Learning Representations (ICLR) 2014*, 2014.

Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual-inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 34(3):314–334, 2015.

Antonio Loquercio, Mattia Segu, and Davide Scaramuzza. A general framework for uncertainty estimation in deep learning. *In 2020 IEEE International Conference on Robotics and Automation (ICRA)*, 5(2):3153–3160, 2020.

David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

Valentin Peretroukhin and Jonathan Kelly. Dpc-net: Deep pose correction for visual localization. *international conference on robotics and automation*, 3(3):2424–2431, 2017.

Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander Alemi, and George Tucker. On variational bounds of mutual information. In *ICML 2019 : Thirty-sixth International Conference on Machine Learning*, pp. 5171–5180, 2019.

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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12240–12249, 2019.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Benoit Steiner, Zachary DeVito, Soumith Chintala, Sam Gross, Adam Paszke, Francisco Massa, Adam Lerer, Gregory Chanan, Zeming Lin, Edward Yang, Alban Desmaison, Alykhan Tejani, Andreas Kopf, James Bradbury, Luca Antiga, Martin Raison, Natalia Gimelshein, Sasank Chilamkurthy, Trevor Killeen, Lu Fang, and Junjie Bai. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS 2019 : Thirty-third Conference on Neural Information Processing Systems*, pp. 8026–8037, 2019.

Jrgen 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*, pp. 573–580, 2012.

Takafumi Taketomi, Hideaki Uchiyama, and Sei Ikeda. Visual slam algorithms: a survey from 2010 to 2016. *IPSJ Transactions on Computer Vision and Applications*, 9(1):16, 2017.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2015.

Naftali Tishby, Fernando C. N. Pereira, and William Bialek. The information bottleneck method. *Proc. 37th Annual Allerton Conference on Communications, Control and Computing, 1999*, pp. 368–377, 2000.

Matias Vera, Pablo Piantanida, and Leonardo Rey Vega. The role of the information bottleneck in representation learning. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 1580–1584, 2018.

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)*, pp. 2043–2050, 2017.

Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks:. *The International Journal of Robotics Research*, 37:513–542, 2018.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *31st Annual Conference on Neural Information Processing Systems, NIPS 2017*, pp. 2524–2533, 2017.

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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8575–8583, 2019.

Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. *arXiv preprint arXiv:2003.01060*, 2020.

Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1983–1992, 2018.

Huangying Zhan, Chamara Saroj Weerasekera, Jiawang Bian, and Ian D. Reid. Visual odometry revisited: What should be learnt? *arXiv preprint arXiv:1909.09803*, 2019.

Jingwei Zhang, Tongliang Liu, and Dacheng Tao. An information-theoretic view for deep learning. *arXiv preprint arXiv:1804.09060*, 2018.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6612–6619, 2017.

A Appendix

A.1 Derivation of the Variational Lower Bound

By the well-established variational bounds for mutual information (MI) (Kingma & Welling, 2014; Alemi et al., 2017; Poole et al., 2019), we directly have a lower bound and a upper bound for the first and second MI in Equation 1, respectively:

$$I(\xi_{1:T}||s_{1:T}) \geq E_{s_{1:T},\xi_{1:T}}[log\ q(\xi_{1:T}|s_{1:T})], \tag{12}$$

$$I(\{o_{1:T}^{(m)}\}_{m=1}^{\mathcal{M}}||s_{1:T}|\xi_{1:T}) \leq E_{\{o_{1:T}^{(m)}\}_{m=1}^{\mathcal{M}},\xi_{1:T}}[D_{KL}[p(s_{1:T}|\{o_{1:T}^{(m)}\}_{m=1}^{\mathcal{M}},\xi_{1:T})||q(s_{1:T}|\xi_{1:T})]]. \tag{13}$$

Also, it is straightforward to show that $E_{x,y}[f(x)] = E_x[f(x)]$ if $f(x)$ is a function that does not depend on y:

$$\begin{aligned} E_{x,y}[f(x)] &= \int_x\int_y p(x,y)f(x)dxdy = \int_x[\int_y p(x)p(y|x)dy]f(x)dx & (14) \\ &= \int_x p(x)[\int_y p(y|x)dy]f(x)dx = \int_x p(x)f(x)dx = E_x[f(x)]. & (15) \end{aligned}$$

Thus, we change the subscripts of the expectations in Equations 12-13 to $s_{1:T}, \xi_{1:T}, \{o_{1:T}^{(m)}\}_{m=1}^{\mathcal{M}}$. For simplicity, we omit the subscripts and denote $\{o_{1:T}^{(m)}\}_{m=1}^{\mathcal{M}}$ as $o_{1:T}$ in the rest of the derivation. We assume Markov property for this sequence processing problem. Then the right-hand side (RHS) of Equation 12 becomes:

$$E[log\ q(\xi_{1:T}|s_{1:T})] = E[log\prod_{t=1}^{T} q(\xi_t|s_t)] = E[\sum_{t=1}^{T} log\ q(\xi_t|s_t)]. \tag{16}$$

The formulation of information bottleneck implies that $\xi \to o \to s$ forms a Markov chain, since the feature encoder for s only depends on the input data o (Tishby et al., 2000; Alemi et al., 2017). Therefore, we have $p(s_{1:T}|o_{1:T},\xi_{1:T}) = p(s_{1:T}|o_{1:T})$. Then by Equation 15 and the Markov assumption, the KL divergence term inside the expectation in the RHS of Equation 13 becomes:

$$\begin{aligned} & D_{KL}[p(s_{1:T}|o_{1:T},\xi_{1:T})||q(s_{1:T}|\xi_{1:T})] & (17) \\ = & \int_{s_{1:T}} p(s_{1:T}|o_{1:T},\xi_{1:T})log\frac{p(s_{1:T}|o_{1:T},\xi_{1:T})}{q(s_{1:T}|\xi_{1:T})}ds_{1:T} & (18) \\ = & \int_{s_{1:T}} p(s_{1:T}|o_{1:T},\xi_{1:T})log\frac{p(s_{1:T}|o_{1:T})}{q(s_{1:T}|\xi_{1:T})}ds_{1:T} & (19) \\ = & \int_{s_{1:T}} p(s_{1:T}|o_{1:T},\xi_{1:T})log\prod_{t=1}^{T}\frac{p(s_t|o_{t-1:t},s_{t-1})}{q(s_t|\xi_t,s_{t-1})}ds_{1:T} & (20) \\ = & \int_{s_{1:T}} p(s_{1:T}|o_{1:T},\xi_{1:T})\sum_{t=1}^{T}log\frac{p(s_t|o_{t-1:t},s_{t-1})}{q(s_t|\xi_t,s_{t-1})}ds_{1:T} & (21) \\ = & \sum_{t=1}^{T} E_{s_{1:T}}[log\frac{p(s_t|o_{t-1:t},s_{t-1})}{q(s_t|\xi_t,s_{t-1})}] = \sum_{t=1}^{T} E_{s_t}[log\frac{p(s_t|o_{t-1:t},s_{t-1})}{q(s_t|\xi_t,s_{t-1})}] & (22) \\ = & \sum_{t=1}^{T} D_{KL}[p(s_t|o_{t-1:t},s_{t-1})||q(s_t|\xi_t,s_{t-1})] & (23) \end{aligned}$$

By sending Equation 16 and Equation 23 to Equation 12 and Equation 13,respectively, we obtain the lower bound of the information bottleneck objective for odometry learning (Equation 2-4).

Remark: All variational IB-based methods origin from Alemi et al. (2017) (Equations 12-13). However, applying IB into a specific domain is non-trivial. The challenge lies in the derivation of proper variational bounds based on specific properties of each problem. This derivation can be more delicate if we incorporate more constraints, potentially from geometric and kinematic insights. Besides, we differ from Dai et al. (2018) and Goyal et al. (2019) in that sequential observations are modeled. From this perspective, our development related to Hafner et al. (2019) and Hafner et al. (2020), from which we further borrowed the motivation of the deterministic component, which by itself is rooted from Chung et al. (2015) and Buesing et al. (2018). Ours differs in that we model the two transition models (Equation 4) separately, each with a deterministic component to improve model capacity (Figure 1(b) and Equations 5-6). Moreover, we theoretically prove that constraining the IB objective essentially upper bounds the expected generalization error and establish the connection between IB and geometry methods, which provides deeper insights into IB-based methods.

A.2 Proof of Lemmas, Theorems, and Corollaries

Lemma 2. *If $X \rightarrow S \rightarrow \xi$ forms a Markov chain and assume $\xi = g(X, \Theta)$ is a one-to-one function w.r.t. X and Θ, then we have*

$$I(X||S) \geq I(X||\xi) = I(X||\Theta) + E_\theta[H(X|\theta)] \geq I(X||\Theta). \tag{24}$$

Proof: By assuming that g is a one-to-one function, we have $p(\xi) = p(x, \theta)$ for an instantiation $g : x, \theta \rightarrow \xi$ for $\xi = g(X, \Theta)$. Then we have:

$$\begin{aligned} I(X||\xi) &= \int_x \int_\xi p(x,\xi) log \frac{p(x,\xi)}{p(x)p(\xi)} dx d\xi \qquad (25)\\ &= \int_x \int_{(x',\theta)} p(x,(x',\theta)) log \frac{p(x,(x',\theta))}{p(x)p(x',\theta)} dx d(x',\theta) \qquad (26)\\ &= \int_{(x,\theta)} p(x,(x,\theta)) log \frac{p(x,(x,\theta))}{p(x)p(x,\theta)} d(x,\theta) \qquad (27)\\ &\quad + \int_x \int_{(x' \neq x,\theta)} p(x,(x',\theta)) log \frac{p(x,(x',\theta))}{p(x)p(x',\theta)} dx d(x',\theta). \qquad (28) \end{aligned}$$

Because $\forall x \neq x',\ p(x,(x',\theta)) = 0$ and $\lim\limits_{a\rightarrow 0} alog(a) = 0$, we have:

$$\int_x \int_{(x' \neq x,\theta)} p(x,(x',\theta)) log \frac{p(x,(x',\theta))}{p(x)p(x',\theta)} dx d(x',\theta) = 0. \tag{29}$$

By $p(x,(x,\theta)) = p(x|x,\theta)p(x,\theta) = p(x,\theta)$, we have:

$$\begin{aligned} & \int_{(x,\theta)} p(x,(x,\theta)) log \frac{p(x,(x,\theta))}{p(x)p(x,\theta)} d(x,\theta) \qquad (30)\\ =\ & \int_{(x,\theta)} p(x,\theta) log \frac{p(x,\theta)}{p(x)p(x,\theta)} d(x,\theta) \qquad (31)\\ =\ & \int_{(x,\theta)} p(x,\theta) log \frac{1}{p(x)} d(x,\theta) = \int_x \int_\theta p(x,\theta) log \frac{1}{p(x)} dx d\theta. \qquad (32) \end{aligned}$$

By combining Equation 29 and Equation 32 with Equation 28, $I(X||\xi)$ becomes:

$$I(X||\xi) = \int_x \int_\theta p(x,\theta) log \frac{1}{p(x)} dx d\theta. \tag{33}$$

Recall the definition of $I(X||\Theta)$:

$$I(X||\Theta) = \int_x \int_\theta p(x,\theta) log \frac{p(x,\theta)}{p(x)p(\theta)} dx d\theta = \int_x \int_\theta p(x,\theta) log \frac{p(x|\theta)}{p(x)} dx d\theta \tag{34}$$

Therefore we have:

$$
\begin{aligned}
I(X||\Theta) - I(X||\xi) &= \int_x \int_\theta p(x,\theta) log(x|\theta) dx d\theta && (35)\\
&= \int_x \int_\theta p(\theta) p(x|\theta) log(x|\theta) dx d\theta && (36)\\
&= -\int_\theta p(\theta)[-\int_x p(x|\theta) log(x|\theta) dx] d\theta && (37)\\
&= -E_\theta[H(x|\theta)] \leq 0 && (38)
\end{aligned}
$$

Because $X \to S \to \xi$ forms a Markov chain, we have $I(X||\xi) \leq I(X||S)$. Then by Equation 38, Lemma 2 holds.

Theorem 3. *Assuming $X \to S \to \xi$ is a Markov chain, the loss function $l(X,\Theta)$ is sub-σ-Gaussian distributed*[2] *and the prediction function $\xi = g(X,\Theta)$ is a one-to-one function w.r.t. the input data and network parameters Θ, we have the following upper bound for the expected generalization error:*

$$E[R(\Theta) - R_T(\Theta)] \leq exp(-\frac{L}{2} log\frac{1}{\eta}) \sqrt{\frac{2\sigma^2}{n} I(X||S)}, \tag{39}$$

where L, η, and n are the effective number of layers causing information loss, a constant smaller than 1, and the sample size, respectively. $R(\Theta) = E_{X\sim D}[l(X,\Theta)]$ is the expected loss value given Θ and $R_T(\Theta) = \frac{1}{n}\sum_{i=1}^{n} l(X_i,\Theta)$ is a sample estimate of $R(\Theta)$ from the training data.

Proof: Assume the loss function $l(X,\Theta)$ is sub-σ-Gaussian distributed, Xu & Raginsky (2017) has proven that the following bound holds for general algorithms with learning parameter set Θ:

$$E[R(\Theta) - R_T(\Theta)] \leq \sqrt{\frac{2\sigma^2}{n} I(X||\Theta)}. \tag{40}$$

Zhang et al. (2018) extended this result to the setting of neural networks and derived the generalization bound for a neural network that has L layers causing information loss:

$$E[R(\Theta) - R_T(\Theta)] \leq exp(-\frac{L}{2} log\frac{1}{\eta}) \sqrt{\frac{2\sigma^2}{n} I(X||\Theta)}, \tag{41}$$

where η is a constant smaller than 1. By Lemma 2 and Equation 41, Theorem 3 holds.

More discussions on Lemma 2 and Theorem 3 The result of Zhang et al. (2018) is interesting in that it provides an explanation for why deeper networks lead to better performance. However, the expected generalization errors in Xu & Raginsky (2017) and Zhang et al. (2018) are both bound by $I(X||\Theta)$, which remains difficult to evaluate in practice. Though their results give a lot of insights into the generalizability of algorithms in information-theoretic language, it is non-trivial to minimize $I(X||\Theta)$ explicitly to control the generalization error bound.

We move one step further by extending their results to $I(X||S)$, the mutual information between input data and latent representations, which itself can be bounded by various well-established variational bounds (Poole et al., 2019) and optimized during training. Our result provides an explanation for the empirical generalization ability of the information bottleneck method, which explicitly minimizes $I(X||S)$. By minimizing $I(X||S)$, we are actually tightening the upper bound of the generalization error, thus leading to better generalization performance.

A related work by Vera et al. (2018) proved a similar result for information bottleneck: "Let $\mathcal{F}$ be a class of encoders. Then, for every P_{XY} and every $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of $\mathcal{S}_n \sim P_{XY}^n$ the following inequality holds $\forall Q_{U|X} \in \mathcal{F}$:

$$\varepsilon_{gap}(Q_{U|X}, \mathcal{S}_n) \leq A_\delta \sqrt{I(\hat{P_X}||Q_{U|X})} \frac{log(n)}{\sqrt{n}} + \frac{C_\delta}{\sqrt{n}} + \mathcal{O}(\frac{log(n)}{n}), \tag{42}$$

where $(A_\delta, B_\delta, C_\delta)$ are quantities independent of the data set $\mathcal{S}_n$: $A_\delta := \frac{\sqrt{2}B_\delta}{P_X(x_{min})}(1 + 1/\sqrt{|X|})$, $B_\delta := 2 + \sqrt{log(\frac{|Y|+3}{\delta})}$ and $C_\delta := 2|U|e^{-1} + B_\delta\sqrt{|Y|}log\frac{|U|}{P_Y(y_{min})}$. $\varepsilon_{gap}(Q_{U|X}, S_n)$

[2] Recall that a random variable l is sub-σ-Gaussian distributed if $E[e^{\lambda(l-E[l])}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$, $\forall\, \lambda \in R$.

is the generalization gap which is defined as $|L_{emp}(Q_{U|X}, \mathcal{S}_n) - -L(Q_{U|X})|$. $L(Q_{U|X})$ and $L_{emp}(Q_{U|X}, \mathcal{S}_n)$ are the true risk and the empirical risks, respectively." We refer readers to Vera et al. (2018) for more details on their result.

Our result differs from that of Vera et al. (2018) in that: (1) Equation 42 only applies to the cross-entropy loss function, while our result holds for a broader range of loss functions under the sub-σ-Gaussian assumption; (2) we provide a tighter generalization bound compared with that of Vera et al. (2018) in terms of sample rate ($\frac{1}{\sqrt{n}}$ vs. $\frac{log(n)}{\sqrt{n}}$); (3) For regression problems and for a large latent space, A_δ and C_δ in Equation 42 could be large due to the positive dependency on $|Y|$ and $|U|$. Besides, $\frac{1}{P_X(x_{min})}$ and $\frac{1}{P_Y(y_{min})}$ might also be large in practice, resulting in a loose bound for the generalization error.

Remark: We now give more discussions on the assumptions of Theorem 3: (1) A Markov chain $X \rightarrow S \rightarrow \xi$ is implicitly implied by neural networks with encoder-decoder structures, since the decoder only takes the encoder output as its input and thus does not depend on X given S. The information bottleneck model, that takes the bottleneck S as the encoder output as well as the decoder input, is essentially encoder-decoder structured. Therefore, the Markov chain assumption on $X \rightarrow S \rightarrow \xi$ also holds for the information bottleneck methods. (2) As discussed in Xu & Raginsky (2017), the sub-σ-Gaussian assumption actually implies a broad range of loss function. For instance, as long as a loss function l is bounded, i.e., $l(\cdot,\cdot) \in [a, b]$, then it is guaranteed to be sub-σ-Gaussian distributed with $\sigma = \frac{b-a}{2}$ (Xu & Raginsky, 2017). The network loss landscape consists of multiple local minima, flat or sharp, and most deep learning methods assume a local Gaussian distribution by using L2 loss (Chaudhari et al., 2017). Sub-σ-Gaussian is more general and provides several superiority over the commonly used Gaussian assumption. Chaudhari et al. (2017) claimed that a flat local minimum is preferred for deep learning optimization algorithms due to the robustness towards parameter perturbations. Sub-σ-Gaussian can well represent such flat local regions, e.g. the almost-flat bounded uniform distribution is sub-σ-Gaussian distributed. It is also worth noting that considering the density of local minima (Chaudhari et al., 2017), σ is not necessarily large for local regions, which can be a concern for the tightness of the generalization bound. Another appealing property is that the sum of sub-σ-Gaussian is still sub-σ-Gaussian, i.e. it can fit a larger region with multiple local minima. (3) The one-to-one function assumption can be conservative due to the complexity of real-world data. For many applications, we may use pretrained models to extract high-level features and use these features as input data. For example, a pretrained FlowNet (Dosovitskiy et al., 2015; Ilg et al., 2017) is usually used in deep odometry learning methods. The input data part of this assumption could arguably hold under such circumstances. Considering the prediction part of this assumption, the cardinality of the space of ξ could be sufficiently large for regression problems and for classification problems, the cardinality of the prediction space could also be large since we usually predict the probabilities of each category. Extending the results to a looser assumption on the network function remains an interesting direction for future research.

Corollary 3. *Given the same assumptions in Theorem 3 and let |S| be the cardinality of the latent representation space, we have*

$$E[R(\Theta) - R_T(\Theta)] \leq exp(-\frac{L}{2}log\frac{1}{\eta})\sqrt{\frac{2\sigma^2}{n}log|S|}. \tag{43}$$

Proof: The relationship between mutual information, entropy and the cardinality of the variable space is well recognized, as given in Cover & Thomas (1991):

$$I(X||S) = H(S) - H(S|X) \leq H(S) \leq log|S|. \tag{44}$$

By Equation 44 and Theorem 3, Corollary 3 holds.

Corollary 4. *Given the same assumptions in Theorem 3 and assume S lies in a d-dimensional subspace of the latent representation space, $sup_{s_i \in S_i} ||s_i|| \leq M, \forall i \in [1, d]$ and S can be approximated by a densely quantized space, the following generalization bound holds:*

$$E[R(\Theta) - R_T(\Theta)] \leq exp(-\frac{L}{2}log\frac{1}{\eta})\sigma\sqrt{\frac{dlog(d)}{n} + 2log(2M)\frac{d}{n} + \frac{d}{n/log(n)}}. \tag{45}$$

Proof: We use the same quantization trick in Xu & Raginsky (2017). We define the covering number $\kappa(r, S)$ as the cardinality of the smallest set $S' \subset S$ $s.t.$ $\forall s \in S,\ \exists s' \in S'$ $with$ $||s - s'|| \leq r$. Assume $sup_{s_i \in S_i}\ ||s_i|| \leq M, \forall i \in [1, d]$ and let $r = 1/\sqrt{n}$, we have $\kappa \leq (2M\sqrt{dn})^d$ (Xu & Raginsky, 2017). We give a proof below for this result, which is omitted in Xu & Raginsky (2017):

We first construct $\tilde{S} \subset S$ that satisfies $\forall s_i \in S_i,\ \exists \tilde{s}_i \in \tilde{S}_i$ $with$ $||s_i - \tilde{s}_i|| \leq \frac{r}{\sqrt{d}}, \forall i \in [1, d]$, where i denotes the dimension of the d-dimensional subspace. Then $\tilde{S}$ also satisfies that $\forall s \in S,\ \exists \tilde{s} \in \tilde{S}$ $with$:

$$||s - \tilde{s}|| = \sqrt{\sum_{i=1}^{d} ||s_i - \tilde{s}_i||^2} \leq \sqrt{\sum_{i=1}^{d} \frac{r^2}{d}} = \sqrt{d\frac{r^2}{d}} = r. \tag{46}$$

For i-th dimension, by the assumption that $sup_{s_i \in S_i}\ ||s_i|| \leq M$, we have $s_i \in [-M, M]$. We can uniformly separate the value range $[-M, M]$ into $\frac{2M}{r/\sqrt{d}}$ intervals. Since each interval has length $\frac{r}{\sqrt{d}}$, we can construct a $\tilde{S}_i$ with cardinality $|\tilde{S}_i| = \frac{2M}{r/\sqrt{d}}$ by including all middle points of the intervals. Let $r = \frac{1}{\sqrt{n}}$, we have $|\tilde{S}_i| = 2M\sqrt{dn}$. We then construct a $\tilde{S}$ by repeating this process for all dimensions. By the denifition of $\kappa(r, S)$ and Equation 46, we have:

$$\kappa(r, S) \leq |\tilde{S}| = \prod_{i=1}^{d} |\tilde{S}_i| = \prod_{i=1}^{d} 2M\sqrt{dn} = (2M\sqrt{dn})^d. \tag{47}$$

When $n \to \infty$, we have $r \to 0$, $S' \to S$, and $\kappa(r, S) \to |S|$. Therefore, by assuming S can be approximated by such a densely quantized space and by Equation 47 and Corollary 3, we have:

$$\begin{aligned} E[R(\Theta) - R_T(\Theta)] &\leq exp(-\frac{L}{2}log\frac{1}{\eta})\sqrt{\frac{2\sigma^2}{n}log(2M\sqrt{dn})^d} && (48) \\ &= exp(-\frac{L}{2}log\frac{1}{\eta})\sigma\sqrt{\frac{dlog(d)}{n} + 2log(2M)\frac{d}{n} + \frac{d}{n/log(n)}}. && (49) \end{aligned}$$

Therefore, Corollary 4 holds.

Theorem 4. *If* $(\{o^{(m)}\}_{m=1}^{\mathcal{M}},\ o^{(\mathcal{M}+1)}) \to S \to \xi$ *forms a Markov chain, then we have,*

$$I(\xi||S) \geq I(\xi||\{o^{(m)}\}_{m=1}^{(\mathcal{M})}) + I(\xi||o^{(\mathcal{M}+1)}|\{o^{(m)}\}_{m=1}^{\mathcal{M}}) - I(o^{(\mathcal{M}+1)}||\{o^{(m)}\}_{m=1}^{\mathcal{M}}|\xi). \tag{50}$$

Proof: From Cover & Thomas (1991), the following two lemmas hold (Lemma 3 and Lemma 4):

Lemma 3. *The inequality for conditional mutual information:*

$$I(X_2||X_1|\xi) \geq I(\xi||X_2|X_1) - I(\xi||X_2) + I(X_1||X_2). \tag{51}$$

Lemma 4. *If* $(X_1, X_2) \to S \to \xi$ *forms a Markov chain, we have:*

$$I(\xi||X_1) + I(\xi||X_2) \leq I(\xi||S) + I(X_1||X_2). \tag{52}$$

By Lemma 3 and Lemma 4, we have:

$$\begin{aligned} I(\xi||S) &\geq I(\xi||X_1) + I(\xi||X_2) - I(X_1||X_2) && (53) \\ &\geq I(\xi||X_1) + I(\xi||X_2|X_1) - I(X_2||X_1|\xi). && (54) \end{aligned}$$

Let X_1 and X_2 denote $\{o^{(m)}\}_{m=1}^{\mathcal{M}}$ and $o^{(\mathcal{M}+1)}$, respectively. Then by Equation 54, Theorem 4 holds.

A.3 Detailed Experimental Settings and More Results

A.3.1 Detailed Network Architecture

The overall network can be separated into four components: **(1) Observation encoders**: For image observation, we first extract the output from the out_conv6_1 layer of a pretrained FlowNet2S (Ilg et al., 2017) model as an intermediate high-level feature, which is then flattened and fed into three MLP layers that have feature size 1024 to obtain image features. Note that the last MLP layer does not use the non-linear activation. For IMU data, we use a two-layer GRU model that has feature size 1024 to extract IMU features; **(2) Deterministic transition models**: For observation-level transition, we first fuse the observation features and concatenate the fused feature with s^o_{t-1} and s^p_{t-1} from last time step. For VINet and InfoVIO, we fuse the features directly by concatenation. For SoftFusion and SoftInfoVIO, we use the same soft fusion strategy proposed in Chen et al. (2019). For HardFusion and HardInfoVIO, we also use the same hard fusion strategy proposed in Chen et al. (2019) while the gumbel temperature linearly degrades from 1 to 0.5 in the first 150 epochs during training and is fixed to 0.5 for testing. For pose-level transition, we tile the 6-DOF poses eight times to a vector of length 48, which is then also concatenated with s^o_{t-1} and s^p_{t-1}. Ground-truth 6-DOF poses are used during training while the predicted poses are used during testing. The concatenated features are then fed into a MLP and a GRU layer to obtain h^o_t and h^p_t, respectively. **(3) Stochastic state estimators**: The deterministic states are fed into two MLP layers to obtain the mean and standard error vectors of the stochastic representation, both with size 128. Note that the last MLP layer does not use the non-linear activation. To avoid a trivial solution, we set the minimum standard error to 0.1 and only predict the residue, where the softplus function is used to guarantee a positive residue. We further use the reparameterization trick proposed in Kingma & Welling (2014) to sample from the stochastic representation distributions, which enables gradient backpropagation through the stochastic representations. **(4) Pose regressor**: We feed the sampled observation-level representation s^o_t into three MLP layers to obtain the translation and rotation prediction results. Both translation and rotation share the first two MLP layers, while we use two separate MLP layers without non-linear activation for translation and rotation, respectively.

All MLP layers with non-linear activation use the Relu function. And all MLP layers except those in observation encoders have feature size 256 and 512 for KITTI and EuRoC, respectively. The state size is set to 128 and 256 for KITTI and EuRoC, respectively. For all baseline models (DeepVO, VINet, SoftFusion, and HardFusion), we remove the pose-level transitions and stochatic state estimators and directly feed h^o_t into the pose regressor for prediction.

A.3.2 Detailed Training and Evaluation Strategies

We used the same training and test splits as Chen et al. (2019). For KITTI, we used sequences 00, 01, 02, 04, 06, 08, and 09 for training and the rest for testing. For EuRoC, we used sequence *MH_04_difficult* for testing and the rest for training. KITTI odometry dataset does not contain synchronized IMU data. Therefore, we manually aligned the 100 Hz IMU records in the raw KITTI data to the 10 Hz image sequences using the corresponding timestamps. EuRoC provides synchronized image and IMU data, collected at 20 Hz and 200 Hz, respectively. Following the practice of previous work (Chen et al., 2019; Clark et al., 2017), we downsampled the image and IMU data in EuRoC to 10 Hz and 100 Hz. By assuming a Gaussian distribution for $q_\theta(\xi_t|s_t)$, we reduced the optimization of Equation 3 to minimizing the L2-norm of the pose errors, resulting in the following loss function:

$$\mathcal{L} = \sum_{n=1}^{N} \alpha||t - \hat{t}|| + \beta||r - \hat{r}||, \tag{55}$$

where t and $\hat{t}$ are the ground-truth and predicted translation. r and $\hat{r}$ are the ground-truth and predicted rotation. We used Euler angles as the quantitative rotation measure. α and β are the translation and rotation error weights, respectively, which were set to 1 and 100 for KITTI and 100 and 20 for EuRoC empirically. We predicted the mean and variance of the stochastic representation s_t and set the minimum variance to be 0.01 to avoid a trivial solution. We set γ in Equation 1 to balance the bottleneck effect. All modeled were trained for 300 epochs using mini-batches of 16 clips containing five frames each. We set an initial learning rate to 1e-4, which was reduced to 1e-5 and 5e-6 at epoch 150 and 250 to stabilize the training process.

We trained and evaluated the odometry model in a clip-wise manner. For evaluation, we use a sliding window strategy s.t. the evaluated clips are overlapped, which means a frame-pair can appear at different positions in a clip. As discussed in Section 5.3, we use a refinement strategy that eliminates the results from the first position and averagely ensembles the rest, which leads to better performance. Following Sturm et al. (2012) and Chen et al. (2019), the averaged root mean squared errors (RMSEs) were used for evaluating both translation and rotation performance.

Remark I: We tested the effectiveness of our method in the supervised odometry learning framework, which requires ground-truth pose labels. In odometry, a good thing is that ground-truth can be obtained from carefully setup sensor suits, which reduces human labors in annotations. Furthermore, two recent research trends may also mitigate this problem, i.e. embodied methods that utilize simulated environments and domain adaptation techniques, and unsupervised learning methods that utilize geometry constraints and train the model jointly with other auxiliary tasks. It is worth noting that our proposed method improves on the representation level and can also be applied in these fields to obtain better latent representations.

Remark II: In odometry learning, we usually use Euler angles or quaternions for rotation representation rather than SO(3) as implied SE(3) due to the redundant parameters in the rotation matrix and the orthogonal constraint. We adopt Euler angles in our experiments and assume a Gaussian distribution in this vector space. Though 3D von Mises-Fisher distribution and 4D-Bingham distribution can be arguably more appropriate to model Euler angles and quaternions respectively, it it non-trivial to evaluate and use them for training in practice.

Remark III: In terms of the choice of hyperparameters like , β, and γ, we basically followed the initial setup of prior works such as Wang et al. (2017); Chen et al. (2019); Hafner et al. (2020) and perform a non-intensive and small-range grid searching. More elegant methods such as relying on the covariance estimates (Peretroukhin & Kelly, 2017) can be considered in future study and applications to new datasets

A.3.3 Comparison between KITTI and EuRoC

KITTI dataset is collected from an autonomous driving car in outdoor scenarios while EuRoC dataset is collected from a MAV in two indoor buildings. Thus these two datasets have different statistics, which may require different network design and training strategy finetuning for each dataset. More specifically, since KITTI is collected during driving, the camera poses mainly contains forward translations and left/right rotations, while for EuRoC, the camera poses from a MAV can have more diverse translation and rotation distribution. As shown in Table 3, since the moving speed of a car is higher than a MAV, the translation scale of KITTI is also larger, while the rotation scale of EuRoC is larger than that of KITTI due to the motion features of MAV.

Table 3: Dataset statistics of KITTI and EuRoC, where the averaged L2-norm values are summarized. x, y, z correspond to the coordinnate system used in KITTI, where x denotes the forward axis, y denotes the upward axis, and z denotes the rightward axis. t and r are the overall L2-norm values for the translation and rotation vectors, respectively.

	$t_x(m)$	$t_y(m)$	$t_z(m)$	$t(m)$	$r_x(^o)$	$r_y(^o)$	$r_z(^o)$	$r(^o)$
KITTI	0.0143	0.0195	0.9666	0.9676	0.1217	0.5381	0.1084	0.6255
EuRoC	0.0388	0.0218	0.0358	0.0660	1.0338	0.8559	0.7403	1.8660

Training a good model for EuRoC is more challenging than for KITTI. The reasons are four-folds: (1) Compared with the similar-looking scenarios in KITTI that mainly contains street views, the scenarios in EuRoC are more diverse, including an intrustrial machine hall and an office room; (2) EuRoC sequences have different difficulty levels by manually adjusted obstacles, which means more carefully designed training strategies such as curriculum learning can be used to improve the performance; (3) The videos collected in EuRoC only contain gray-scale images while those in KITTI contain RGB images instead. Considering the FlowNet model was pretrained using RGB images, the domain gap for using gray-scale images should also be taken into account for better performance; (4) The translation scale of EuRoC is much smaller, which can cause difficulty for accurate prediction.

A.3.4 More Ablation Experiments

Effect of the deterministic component: We conducted stochastic-only ablation experiments to examine the effect of the deterministic component in Equation 5-6 by removing the deterministic nodes in Figure 1(b). We implemented two versions depending on whether the observation- and pose-level latent representations (s^o and s^p) were both used as the recurrent network state (StochasticVO/VIO-d), or not (StochasticVO/VIO-s). Results are given in Table 4. Without the deterministic component, both translation and rotation performance dropped significantly, which supports the effectiveness of the deterministic component.

Remark: For stochastic-only models, we remove the stochastic state estimators and let the GRU layer in the deterministic transition models directly output the means and standard error residues of the stochastic representation. For state transitions, we then used sampled states as the transitioned state context for the transition model at next time step. We give more details of the two implementations below. StochasticVO/VIO-d is short for "stochastic VO/VIO with double transition states", which used (s^o_{t-1}, s^p_{t-1}) as the transition state from last time step for both observation- and pose-level transitions. StochasticVO/VIO-s is short for "stochastic VO/VIO with single transition states", which used (s^o_{t-1}, s^o_{t-1}) and (s^p_{t-1}, s^p_{t-1}) as the transition state from last time step for observation- and pose-level transitions, respectively.

Table 4: Results of the stochastic-only models on KITTI.

Model	$t(m)$	$r(^o)$
StochasticVO-s	0.0758	0.0931
StochasticVO-d	0.0783	0.0899
InfoVO (full)	0.0607	0.0869
StochasticVIO-s	0.0714	0.0512
StochasticVIO-d	0.0734	0.0507
InfoVIO (full)	0.0580	0.0416

Effect of the sample size: We study the effect of the sample size by using different ratios r_n of training samples for training the model. Recall that we let the minimum variance to be 0.01 to avoid a trivial solution, which sets a empirical lower bound of the uncertainty. Table 5 shows that a larger sample size reduces both the uncertainty and prediction errors. An interesting observation from our results is though more training samples still benefit the prediction performance, the average variance or the uncertainty measure does not reduce after half of the dataset is added. We suspect that this may be due to the fact that KITTI sequences exhibit quite similar patterns (mostly road driving scenarios). Thus half samples are sufficient for the model to be "familiar" with the dataset and reach the uncertainty margin. While if the training samples are not sufficient enough, e.g. $1/4$ of total samples, the variance increases significantly.

Table 5: Results of varied sample sizes on KITTI. r_n: the ratio of training samples. $\bar{\sigma}^2$: the averaged variance of the latent representation.

r_n	$t(m)$	$r(^o)$	$\bar{\sigma}^2$
$1/4$	0.1977	0.1040	0.0109
$1/2$	0.0602	0.0644	0.0101
$3/4$	0.0589	0.0544	0.0102
$full$	0.0580	0.0416	0.0102

Effect of extra sensors: Motivated by Theorem 4 and our failure-awareness analysis, we study the performance gain of IMU given images and vice versa. The comparison between InfoVO and InfoVIO provides the performance gain of IMU given images. Similarly, to study the performance gain of images given IMU, We trained an IMU-only model, denoted as InfoIO, which is then compared with InfoVIO. The results are summarized in Table 6, which implying that IMU is more 'useful' than cameras for rotation prediction while cameras are more crucial than IMU for translation prediction.

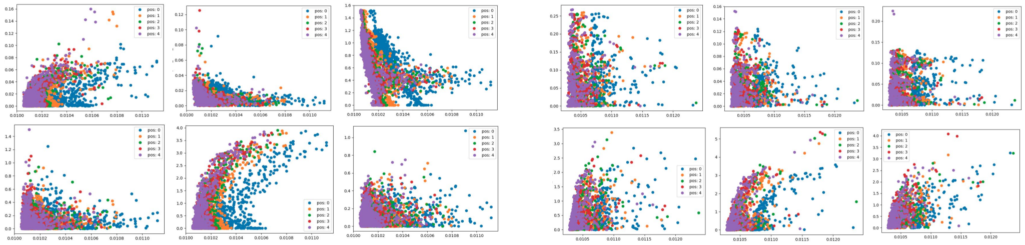
Moreover, IMU provides a larger performance gain in EuRoC than KITTI, which is consistent with fact that the synchronization in EuRoC between IMU and ground-truth poses are more accurate. We also observed that InfoIO performs poorly in KITTI. The large performance gain of images given IMU in KITTI w.r.t. both translation and rotation might also result from the inaccurate alignment of IMU records from the raw KITTI dataset to the image and ground-truth pose sequences.

Table 6: Performance gain of IMU given images and images given IMU.

Model	KITTI		EuRoC	
	$t(m)$	$r(^o)$	$t(m)$	$r(^o)$
InfoIO	0.2069	0.1164	0.0667	0.0740
InfoVO	0.0607	0.0869	0.0310	0.2061
InfoVIO	0.0580	0.0416	0.0276	0.0744

A.3.5 More Uncertainty Results

We show the full uncertainty results of InfoVIO for KITTI and EuRoC in Figure 4. Since the translations along x and y axes and the rotations around x and z axes are relatively small in KITTI dataset, their uncertainties do not show a clear pattern. While for the translation along the forward axis and the rotation around the upward axis (turning left/right), a clear negative and a clear positive relationship is observed for each motion. The reason for this can be that a large forward parallax provides more distinctive matching features for pose prediction while a large turning angle instead dramatically reduces the shared visible areas and results in difficulties to achieve accurate prediction. For the EuRoC dataset, we observed a consistent positive relationship for all three rotations, which makes sense in that the MAV rotations are more uniformly distributed along the three axes. The negative relationship in the translation results of EuRoC is more obscure than that of KITTI, partly due to the difficulties in accurately predicting MAV translations, as discussed in A.3.3.



(a) Uncertainty Results for KITTI

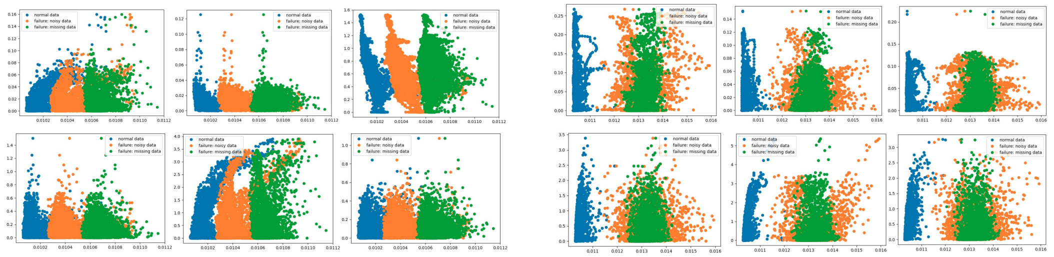
(b) Uncertainty Results for EuRoC

Figure 4: Full uncertainty results of InfoVIO for (a) KITTI and (b) EuRoC. The top and bottom rows represent translation and rotation results. The first, second, and third columns represent x, y, and z, respectively. x, y, z are with respect to the coordinate system in KITTI. pos-i means the result is evaluated at the i-th position in a clip.

Remark: There is also a line of work that attempts to combine learning based methods with geometry based pipelines (Peretroukhin & Kelly, 2017; Yang et al., 2020), where uncertainty plays an important role by serving as a quality measure to properly weigh the learned results. The recent successful work by Yang et al. (2020) used learned aleatoric uncertainty to integrate learned results into the DVO pipeline and achieves SOTA performance in monocular odometry. Our work makes contribution in that we do not explicitly learn the variance of final prediction, but use the variance of the intrinsic latent state instead as the uncertainty measure, which we empirically show that can capture the epistemic uncertainty as well and holds the potential to provide a better fusion guidance. It remains an interesting future research direction to see whether our uncertainty measure can really benefit this hybrid pipeline that combines the merits of both learning and geometry methods.

A.3.6 Detailed Settings and Results for Failure-Awareness Experiments

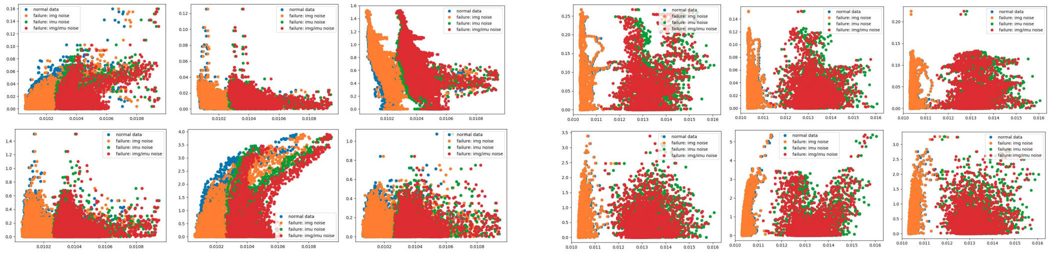
We considered two failure cases, namely, degradations with noisy data and missing data. We add Gaussian noise with mean 0 and standard error 0.1 to the observations in test dataset to create noisy data. For missing data, we replace the observations by this Gaussian noise.



(a) Uncertainty Results with Noisy/Missing Data for KITTI

(b) Uncertainty Results with Noisy/Missing Data for EuRoC

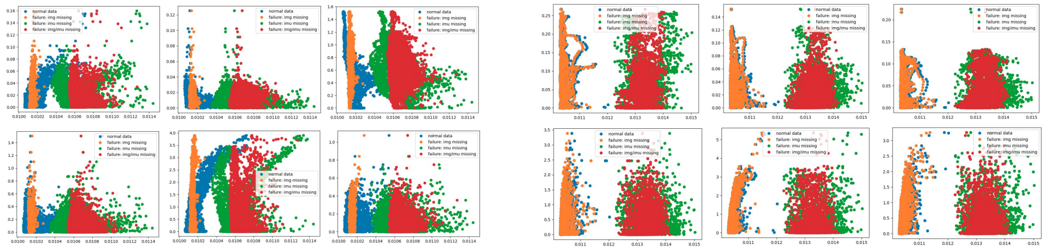
Figure 5: Uncertainty results of InfoVIO on both noisy and missing data for (a) KITTI and (b) EuRoC. The top and bottom rows represent translation and rotation results. The first, second, and third columns represent x, y, and z, respectively. x, y, z are with respect to the coordinate system in KITTI. Blue, orange, and green circles denote results from normal data, noisy data, and missi2ng data, respectively. Both images and IMU records were degraded.



(a) Uncertainty Results with Noisy Data for KITTI

(b) Uncertainty Results with Noisy Data for EuRoC

Figure 6: Uncertainty results of InfoVIO on noisy data for (a) KITTI and (b) EuRoC. The top and bottom rows represent translation and rotation results. The first, second, and third columns represent x, y, and z, respectively. x, y, z are with respect to the coordinate system in KITTI. Blue, orange, green, and red circles denote results from normal data and degraded data with images, IMU, and both images and IMU being noisy, respectively.



(a) Uncertainty Results with Missing Data for KITTI

(b) Uncertainty Results with Missing Data for EuRoC

Figure 7: Uncertainty results of InfoVIO on missing data for (a) KITTI and (b) EuRoC. The top and bottom rows represent translation and rotation results. The first, second, and third columns represent x, y, and z, respectively. x, y, z are with respect to the coordinate system in KITTI. Blue, orange, green, and red circles denote results from normal data and degraded data with images, IMU, and both images and IMU missing, respectively.

The results are shown in Figure 5, Figure 6, and Figure 7. Our models becomes more uncertain as the data degrades. The uncertainty reaches the highest when the data is missing, as expected. A more interesting observation is that the quality of IMU data dominates the uncertainty for both KITTI and EuRoC, implying that current image encoders are not trained well enough and a better image encoder is desirable to fully utilize visual information. Also, data degradation on IMU records leads to higher uncertainty in EuRoC than in KITTI. We suspect this is because the synchronization between the ground-truth poses and IMU records are more accurate in EuRoC than in KITTI. These observations support that the intrinsic uncertainty measure provides a practical tool for failure diagnosis, such as noises, sensor malfunctions, and even mis-synchronization between sensors.