International Journal of Computer Vision (2022) 130:2553–2570
https://doi.org/10.1007/s11263-022-01659-9

Information-Theoretic Odometry Learning

Sen Zhang[1] · Jing Zhang[1] · Dacheng Tao[1]

Received: 18 February 2021 / Accepted: 21 July 2022 / Published online: 12 August 2022

Abstract
In this paper, we propose a unified information theoretic framework for learning-motivated methods aimed at odometry estimation, a crucial component of many robotics and vision tasks such as navigation and virtual reality where relative camera 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.

Keywords Odometry learning · Simultaneous localization and mapping · Information bottleneck · Generalization bound

1 Introduction

Odometry aims to predict six degrees of freedom (6-DOF) relative camera 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; Zhang & Tao, 2020). 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. The relative camera pose is recovered using geometric clues and motion models. Classic geometric methods usually formulate the odometry problem as an optimization problem by incorporating well-established geometric and motion constraints as the objective functions. Nevertheless, due to the complexity and diversity of real-world environments, the explicitly modeled constraints can hardly explain all aspects of the sensor data. Though successful in some real-world scenarios, geometric systems fail to work when the underlying assumptions behind the optimization objectives, such as static environments, discriminative visual features, noiseless observations and brightness constancy, are violated in the real world. Furthermore, since odometry is essentially a time-series prediction problem, how to properly handle time dependency and environment dynamics presents further challenges. Classic geometric methods use filtering or bundle adjustments to take the temporal information into account, while the implicitly implied error distributions might not hold in practice.

Recently end-to-end deep learning methods provide an alternative solution for the odometry problem, which relieves the above-mentioned intrinsic problems in geometric methods. Learning-based methods tackle this problem from another perspective that does not explicitly model the constraints for optimization but learns the mapping from sensor data to camera pose implicitly from large-scale datasets (Wang et al., 2017; Clark et al., 2017; Xue et al., 2019). It has been shown that well-trained deep networks are able to effectively capture the inherent complexity and diversity of the training data and establish the mapping between visual/sequential inputs to desired targets in many computer vision tasks (He et al., 2016; Xu et al., 2021; Zhang et

Communicated by A. Hilton.

✉ Jing Zhang
jing.zhang1@sydney.edu.au; chaimi.ustc@gmail.com

Sen Zhang
szha2609@uni.sydney.edu.au

Dacheng Tao
dacheng.tao@sydney.edu.au

[1] The University of Sydney, Sydney, NSW, Australia

al., 2022a), thus holding promise for addressing the limitations of geometric approaches. In addition, learning-based frameworks can implicitly learn calibrated representations and require no explicit calibration procedures. For monocular visual odometry, the absolute scale can also be recovered from training data, which instead is a non-trivial challenge for geometric methods.

Although existing deep odometry learning methods have performed competitively against their geometric counterparts, they still fail to satisfy some basic requirements. First of all, 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. In addition, 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 the potential 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 the observations and relative camera 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 the 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 expected 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 expected generalization error, providing a theoretical explanation for the complexity-overfitting trade-off in the latent representation space. When the 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 relative camera 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 geometric 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;
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 geometric methods (Zhan et al., 2020; Yang et al., 2020; Zhang et al., 2022c, b). 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 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 deterministic-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 deterministic 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 and Raginsky (2017) in information-theoretic language. This work was subsequently extended by Zhang et al. (2021) 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 $\{o_{t-1:t}^{(m)}\}_{m=1}^{\mathcal{M}}$ from $\mathcal{M}$ sensors (e.g. camera, IMU and lidar), where t is the time index. This pose prediction problem can be formulated as $\xi_t = g(\{o_{t-1:t}^{(m)}\}_{m=1}^{\mathcal{M}}, \Theta)$, 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 1a shows a typical procedure shared by representative deep odometry learning methods.

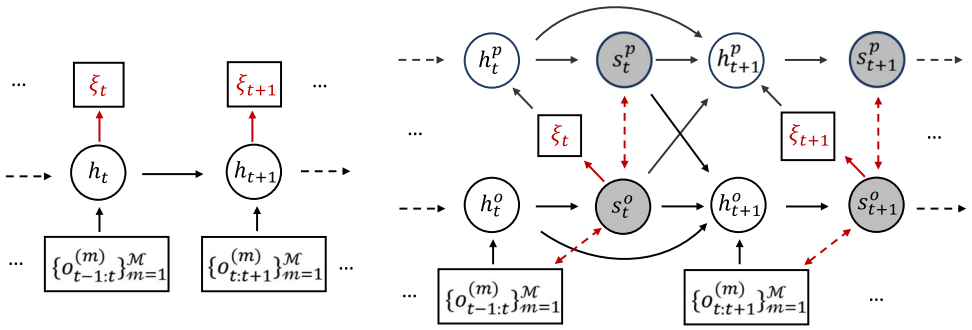
In many settings, observations are of high dimensionalities, such as images and lidar 3D points. Geometric methods use low-dimensional features to represent observations, while learning-based methods learn a representation from training data. However, both features may contain pose-irrelevant information that is 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 the 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 $I(\{o_{1:T}^{(m)}\}_{m=1}^{\mathcal{M}}||s_{1:T}|\xi_{1:T})$ 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 $\{o_{1:T}^{(m)}\}_{m=1}^{\mathcal{M}}$ and pose annotations $\xi_{1:T}$ from time 1 to T, our information-theoretic odometry learning problem is:

$$max_{\Theta}\ \mathcal{J}(\Theta) = I(\xi_{1:T}||s_{1:T}) - \gamma I_{\text{bottleneck}}, \tag{1}$$

$$I_{\text{bottleneck}} = I(\{o_{1:T}^{(m)}\}_{m=1}^{\mathcal{M}}||s_{1:T}|\xi_{1:T}), \tag{2}$$

where the IB weight γ controls the trade-off between the two MI terms. By Eq. (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:



(a) Classic Odometry Learning **(b)** Deterministic-Stochastic IB Odometry Learning

Fig. 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 (Color figure online)

$$\mathcal{J}(\Theta) \geq \mathcal{J}'(\Theta) = E_{s_{1:T},\{o_{1:T}^{(m)}\}_{m=1}^{\mathcal{M}},\xi_{1:T}}[\sum\nolimits_{t=1}^{T} J_t'], \tag{3}$$

$$J_t' = \mathcal{J}_t^{pose} - \gamma \mathcal{J}_t^{\text{bottleneck}}, \tag{4}$$

$$\mathcal{J}_t^{pose} = log\ q_\theta(\xi_t|s_t), \tag{5}$$

$$\mathcal{J}_t^{\text{bottleneck}} = D_{KL}(p_\phi||q_\varphi)., \tag{6}$$

$$p_\phi = p_\phi(s_t|\{o_{t-1:t}^{(m)}\}_{m=1}^{\mathcal{M}}, s_{t-1}), \tag{7}$$

$$q_\varphi = q_\varphi(s_t|\xi_t, s_{t-1}). \tag{8}$$

The detailed derivation is provided in the Supplementary Material. This lower bound consists of a variational pose regressor $q_\theta(\xi_t|s_t)$, an observation-level transition model $p_\phi(s_t|\{o_{t-1:t}^{(m)}\}_{m=1}^{\mathcal{M}}, s_{t-1})$, and a pose-level transition model $q_\varphi(s_t|\xi_t, s_{t-1})$, all of which are modeled by neural networks. For simplicity, we denote the representations from the observation-level and pose-level transition models s_t^o and s_t^p, respectively. In practice, s_t^o is used for the pose regressor. Intuitively, minimizing the KL divergence in Eq. (6) forces the distribution of s_t^o to approximate that of s_t^p which does not encode the observation information at time t, thus regularizing s_t^o 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 Eq. (6) as:

$$\textbf{observation-level}: \ p_\phi(s_t^o|h_t^o), \tag{9}$$

$$h_t^o = f^o(h_{t-1}^o, \{o_{t-1:t}^{(m)}\}_{m=1}^{\mathcal{M}}, s_{t-1}^o, s_{t-1}^p), \tag{10}$$

$$\textbf{pose-level}: \ q_\varphi(s_t^p|h_t^p), \tag{11}$$

$$h_t^p = f^p(h_{t-1}^p, \xi_t, s_{t-1}^o, s_{t-1}^p). \tag{12}$$

We use two deterministic functions f^o and f^p for observation- and pose-level transitions, respectively, which are both modeled by recurrent neural networks. In addition, both s_{t-1}^o and s_{t-1}^p are used for the two deterministic transition functions to help to reduce the KL divergence between the distributions of s_t^o and s_t^p. 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_t^p. Fig. 1b shows the overall framework of our method.

Remark I. 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 Sect. 5.3. Of note is that it is straightforward 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.

Remark II. All variational IB-based methods origin from Alemi et al. (2017). However, applying IB into a specific domain is non-trivial. The challenge lies in the derivation of proper variational bounds based on the 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 [Eq. (6)] separately, each with a deterministic component to improve model capacity

(Fig. 1b and Eqs. (9)–(11)). Moreover, we theoretically prove that constraining the IB objective essentially upper bounds the expected generalization error and establish the connection between IB and geometric methods in Sect. 4, which provides deeper insights into IB-based methods.

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., 2021). 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 geometric methods is also established to provide further insights.

4.1 Generalization Bound for Information Bottleneck

Xu and Raginsky (2017) and Zhang et al. (2021) 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 1 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*

$$I(X, S) \geq I(X, \xi) = I(X, \Theta) + E_\theta[H(X|\theta)] \quad (13)$$

$$\geq I(X, \Theta). \quad (14)$$

Lemma 1 enables us to extend the generalizability results for neural networks regarding $I(X, \Theta)$ (Zhang et al., 2021) 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:*

$$E[R(\Theta) - R_T(\Theta)] \leq \exp(-\frac{L}{2} log \frac{1}{\eta}) \sqrt{\frac{2\sigma^2}{n} I(X, S)}, \quad (15)$$

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 works is that we bound the generalization error by $I(X, S)$ which is minimized in Eq. (1) rather than $I(X, \Theta)$ which is hard to evaluate. By Theorem 1, 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 that 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 shown in Xu and Raginsky (2017) and Zhang et al. (2021). The detailed proof of Lemma 1 and Theorem 1 can be found in the Supplementary Material.

Remark I. The result of Zhang et al. (2021) is interesting in that it provides an explanation for why deeper networks lead to better performance. However, the expected generalization errors in Zhang et al. (2021) and Xu and Raginsky (2017) are both bounded 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 IB method, which explicitly minimizes $I(X||S)$. By minimizing $I(X||S)$, we are actually tightening

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

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 IB: "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 $\forall Q_{U|X} \in \mathcal{F}$:

$$\varepsilon_{gap}(Q_{U|X}, \mathcal{S}_n) \leq A_\delta \sqrt{I(\hat{P}_X||Q_{U|X})} \frac{log(n)}{\sqrt{n}} + \frac{C_\delta}{\sqrt{n}} + \mathcal{O}\left(\frac{log(n)}{n}\right), \tag{16}$$

where $(A_\delta, B_\delta, C_\delta)$ are quantities independent of the data set $\mathcal{S}_n$: $A_\delta := \frac{\sqrt{2}B_\delta}{P_X(x_{min})}(1 + 1/\sqrt{|X|})$, $B_\delta := 2 + \sqrt{log\left(\frac{|Y|+3}{\delta}\right)}$ and $C_\delta := 2|U|e^{-1} + B_\delta\sqrt{|Y|}log\frac{|U|}{P_Y(y_{min})}$. $\varepsilon_{gap}(Q_{U|X}, S_n)$ 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) Eq. 16 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) w.r.t. 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 Eq. (16) 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 II. We now give more discussions on the assumptions of Theorem 1: (1) a Markov chain $X \to S \to \xi$ is implicitly implied in 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. In this case, we have $P(\xi|S) = P(\xi|S, X)$. It is worth noting that in more general settings where more flexible network structures that allow additional connections between X and ξ are used, this Markov chain assumption may not hold. However, for the IB methods, since an IB model is essentially encoder-decoder structured by constraining the information flow between the encoder and the decoder, the Markov chain assumption on $X \to S \to \xi$ holds under this setting. (2) As discussed in Xu and Raginsky (2017), the sub-σ-Gaussian assumption actually implies a broad range of loss functions. 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 superiorities 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.

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 1 and let $|S|$ be the cardinality of the latent representation space, we have*

$$E[R(\Theta) - R_T(\Theta)] \leq \exp\left(-\frac{L}{2}log\frac{1}{\eta}\right)\sqrt{\frac{2\sigma^2}{n}log|S|}. \tag{17}$$

It is well recognized that a large model complexity can impair the generalizability of the model. We reveal this complexity-overfitting trade-off in Corollary 1, where the expected generalization error is upper bounded by the cardinality of the latent representation space. In addition, considering the model design and sample collection, Corollary 1 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 1 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:

$$E[R(\Theta) - R_T(\Theta)] \leq exp\left(-\frac{L}{2}log\frac{1}{\eta}\right)\sigma\mathcal{C}, \tag{18}$$

$$\mathcal{C} = \sqrt{\frac{dlog(d)}{n} + 2log(2M)\frac{d}{n} + \frac{d}{n/log(n)}}. \tag{19}$$

In practice, it is usually difficult to evaluate $log|S|$ in Corollary 1 numerically. Therefore, we leverage the quantization trick used in Xu and 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 2, 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 $X \rightarrow S \rightarrow \xi$.

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

$$I(\xi||S) \geq I_{old} + I_{new} - I_{obs}, \tag{20}$$

$$I_{old} = I(\xi||\{o^{(m)}\}_{m=1}^{(\mathcal{M})}), \tag{21}$$

$$I_{new} = I(\xi||o^{(\mathcal{M}+1)}|\{o^{(m)}\}_{m=1}^{\mathcal{M}}), \tag{22}$$

$$I_{obs} = I(o^{(\mathcal{M}+1)}||\{o^{(m)}\}_{m=1}^{\mathcal{M}}|\xi). \tag{23}$$

Theorem 2 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. In addition, we further observe that 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 Geometric Methods

More generally, an odometry system can be modeled as $h(z_{k,j}, v_k, \check{x_k}) \rightarrow (\hat{x_k}, p_j)$ 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 $\check{x_k}$. In addition, 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 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 further examine the ability of generalization to more challenging scenarios such as extreme weather and lighting conditions by testing DeepVO and InfoVO on vKITTI2 (Cabon et al., 2020). In addition, we empirically study the pose-irrelevant information contained in DeepVO and InfoVO to examine the underlying hypothesis of the problem that we target. We also conducted extensive ablation studies on the deterministic component, the weight of the IB objective, the sample size, extra sensors, the intrinsic uncertainty measure, and the growing rate relationship between the latent dimension and $n/log(n)$.

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). More advanced optical flow estimation methods could also be explored such as RAFT Teed and Deng (2020) and GMFlow Xu et al. (2022). The other parts were modeled by MLP layers.

5.1.1 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 the 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. The features are concatenated in VINet and InfoVIO. For SoftFusion, SoftInfoVIO, HardFusion and HardInfoVIO, we also use the same soft and 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. We tile the 6-DOF poses eight times to a vector of length 48 for the pose-level transition, 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 an 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 and 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 have feature sizes 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 stochastic state estimators and directly feed h^o_t into the pose regressor for prediction.

5.1.2 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 the 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, respectively. By assuming a Gaussian distribution for $q_\theta(\xi_t|s_t)$, we reduced the optimization of Eq. (3) to minimizing the L2-norm of the pose errors, resulting in the following loss function:

$$\mathcal{L} = \sum_{n=1}^{N} \alpha||t - \hat{t}|| + \beta||r - \hat{r}||, \tag{24}$$

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 Eq. (1) to balance the bottleneck effect. All models 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 used a sliding window strategy s.t. the evaluated clips are overlapped, which means a frame-pair can appear at different positions in a clip. A refine-

ment strategy that eliminates the results from the first position and averagely ensembles the rest was designed based on our empirical observations, which will be discussed in Sect. 5.3. 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. In odometry learning, we usually use Euler angles or quaternions for rotation representation rather than SO(3) as implied in SE(3) due to the redundant parameters of the rotation matrix and the orthogonal constraint. In this work, we adopt Euler angles in our experiments and assume a Gaussian distribution in this vector space for simplicity and easier implementation. Though 3D von Mises–Fisher distribution Khatri and Mardia (1977) and 4D-Bingham distribution Gilitschenski et al. (2019) can be arguably more appropriate to model Euler angles and quaternions, respectively, it is non-trivial to evaluate and use them for training in practice. The exploration of these more advanced representation and distribution choices remains potentially important future research work.

Remark II. In terms of the choice of hyperparameters like α, 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 performed 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 studies and applications to new datasets.

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 (Mourikis & Roumeliotis, 2007; 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 averaged 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. EuRoC by its nature is much more challenging than KITTI. The major difficulties include (1) the diverse scenarios including an industrial machine hall and an office room, compared with the similar-looking street views in KITTI, (2) the varying difficulty levels of different sequences by manually adjusted obstacles, and (3) the grey-scale images while the FlowNet encoder was pretrained using RGB images, which indicates a domain gap from RGB to grey images and thus degrades the results accordingly. Therefore, reducing the performance gap on EuRoC may require more carefully designed training strategies. Comparisons between the two datasets are summarized in the Supplementary Material.

Table 1 Test results on KITTI and EuRoC

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

We report the averaged RMSEs for translation and rotation, respectively. The best results for VO and VIO are shown in bold

†Results of MSCKF on KITTI and OKVIS on EuRoC are from Chen et al. (2019)

5.2.1 Visualization of KITTI Trajectories

We further provide per sequence result and trajectory visualization for DeepVO, InfoVO, VINet and InfoVIO to illustrate the benefit of optimizing the IB objective.

Results of the test sequences 05, 07, and 10 are presented in Table 2 and Fig. 2. Though long-term accumulated drifts are observed for all end-to-end learning-based odometry meth-

Table 2 Per sequence results on KITTI. We report the averaged translation RMSE drift t_{rel} (%) on length of 100–800 m and the averaged rotation RMSE drift r_{rel} (^{o}/100 m) on length of 100–800 m

Model	05		07		10	
	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
DeepVO	6.25	2.29	5.66	3.60	7.12	**1.91**
InfoVO	**4.30**	**1.54**	**4.52**	**3.34**	**6.25**	2.16
VINet	3.52	1.08	5.39	3.43	8.58	2.89
InfoVIO	**3.33**	**0.91**	**4.69**	**3.00**	**7.43**	**2.44**

The best results for VO and VIO are shown in bold

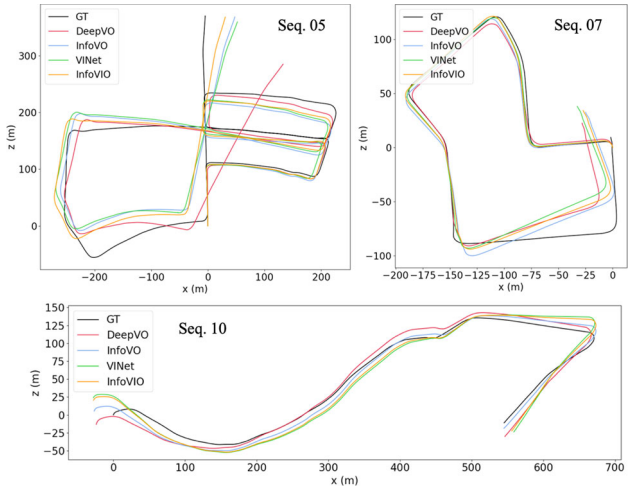


Fig. 2 Predicted trajectories of DeepVO, InfoVO, VINet, and InfoVIO on KITTI sequences 05, 07 and 10 (Color figure online)

Table 3 Results on challenging sequences on vKITTI2. W and L denotes sequences that contain different weather conditions (rain and fog) and lighting conditions (morning, sunset, and overcast), respectively

Model	Conditions	$t(m)$	$r(^o)$
DeepVO	W	1.5214	0.1676
InfoVO	W	1.5011	0.1368
DeepVO	L	1.4642	0.1524
InfoVO	L	1.3614	0.1239

ods, InfoVO and InfoVIO that optimize the IB objective still perform better than DeepVO and VINet, especially on sequence 05, which is longer and more challenging due to the increased number of turns.

5.3 Generalization to Challenging Scenarios

In addition to the results reported on the test splits of KITTI and EuRoC, we further examine the performance of InfoVO on vKITTI2 (Cabon et al., 2020), a simulated autonomous driving dataset that contains various scenarios. We illustrate the benefit of the IB objective by training DeepVO and InfoVO on the clean sequences in vKITTI2 and comparing their performance on the more challenging counterparts that have different weather conditions (rain and fog) and lighting conditions (morning, sunset, and overcast). We used Scene 01, 02, and 06 as the training set and left Scene 18 and 20 as the test set. Of note is that only the clean sequences in the training set are used during training.

Results under different weather and lighting conditions are presented in Table 3. It is shown that InfoVO achieves better generalization results in the challenging scenarios than DeepVO w.r.t. both translation and rotation predictions. Besides, our results suggest extreme weather conditions present more challenging than different lighting conditions due to the noises and texture losses in the frames, which remains an interesting research direction towards a more robust odometry system in those challenging scenarios.

5.4 Compactness of the Latent Space

A key hypothesis underlying the motivation to develop our framework is that methods without specific consideration on the compactness of the latent space will implicitly encode pose-irrelevant information into the learnt features, which can be eliminated by the information bottleneck objective. We empirically demonstrated this phenomenon by comparing the reconstruction accuracies using the features learnt by DeepVO and InfoVO.

Since the optical flow features from the pretrained FlowNet2S (Ilg et al., 2017) are used as the network inputs for both DeepVO and InfoVO, we proposed to empirically measure the amount of pose-irrelevant information by the ability to reconstruct those optical flow features from the latent space of DeepVO and InfoVO, respectively. Specifically, we used three MLP layers as the reconstruction decoder, which takes the latent features from the DeepVO and InfoVO models trained on the KITTI dataset as input. We varied the hidden size d of the decoder to examine the performance under different reconstruction capacities. We adopted the same training/test split as in our main experiment and trained the decoder for 300 epochs.

The results of the averaged MSE loss $\bar{l}$ for optical flow feature reconstruction using different hidden sizes are presented in Table 4. We also reported the results by taking white Gaussian noise as input. The input optical flow vectors contain both pose-relevant and pose-irrelevant information, such as occlusions and the motion of dynamic objects. Since

Table 4 Results of the reconstruction of optical flow features on KITTI

Model	d	$\bar{l}$
DeepVO	1024	0.0387
DeepVO	512	0.0391
DeepVO	256	0.0396
DeepVO	128	0.0401
InfoVO	1024	0.0444
InfoVO	512	0.0456
InfoVO	256	0.0508
InfoVO	128	0.0530
Noise $\sim N(0,1)$	1024	0.0541
Noise $\sim N(0,1)$	512	0.0541
Noise $\sim N(0,1)$	256	0.0541
Noise $\sim N(0,1)$	128	0.0541

InfoVO achieves a higher accuracy than DeepVO in terms of pose prediction, which indicates that InfoVO has extracted more pose-relevant information than DeepVO to achieve this, the inferiority of InfoVO to reconstruct optical flow features indicates that InfoVO has eliminated more pose-irrelevant information than DeepVO, while maintaining pose-relevant information from the optical flow features for downstream pose prediction tasks. It is worth noting that the reconstruction performance of InfoVO is close to that of random noise using the hidden size 128, which means although a certain degree of pose-irrelevant information may still exist in the feature space of InfoVO, the remaining amount is small, and it requires a relatively powerful decoder to extract this information.

5.5 Growing Rate of the Latent Dimension

As suggested in Corollary 2, the growing rate of the latent dimension d should not exceed that of $n/\log(n)$ to avoid overfitting and achieve a tighter generalization bound. To illustrate this effect, we use different sample size ratios for sequence 01 to train InfoVO, and test the trained models on sequences 09 and 10 that have quite different motion patterns (slower vehicle speed) with sequence 01. We first choose the sample size ratio $r_0 = 1/4$ as the starting point, and empirically determine its corresponding latent dimension $d_0 = 384$ that leads to neither underfitting nor overfitting. Then we study the performance of InfoVO models using different latent dimensions under the sample size ratios $r_1 = 1/2$ and $r_2 = 1.0$, whose growing rates of $n/log(n)$ are 1.780 and 3.208, respectively. The results are presented in Fig. 3.

We examine the results of latent dimensions 256, 512, 1024, 1536, and 2048. For $r_1 = 1/2$ and $r_2 = 1.0$, the latent dimensions that have the same growing rates as $n/log(n)$ are $384 * 1.780 \approx 684$ and $384 * 3.208 \approx 1232$, respectively.

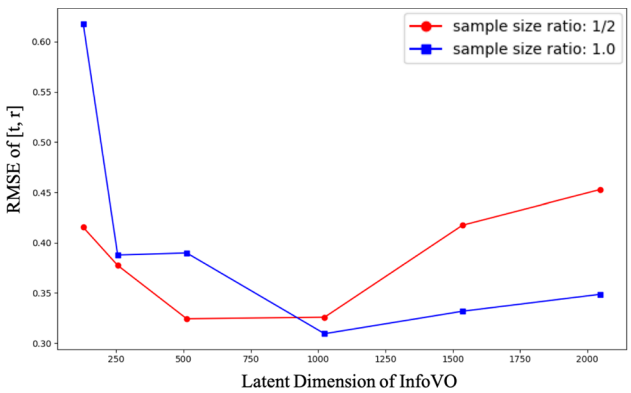


Fig. 3 Results of varying latent dimensions (256, 512, 1024, 1536, 2048) under the sample size ratios 1/2 (red) and 1.0 (blue). The RMSE results of the combined 6-DOF translation and rotation vector are reported

Accordingly, our results showed that the latent dimensions 512 and 1024 achieved the best test results before overfitting for $r_1 = 1/2$ and $r_2 = 1.0$, respectively. A small latent dimension led to an underfitted model while overfitting was observed when the growing rate of the latent dimension exceeds that of $n/log(n)$, which supports Corollary 2 empirically.

5.6 Ablation Studies

Extensive ablation studies were conducted to examine the effects of (1) the deterministic component, (2) the IB weight, (3) the sample size and (4) extra sensors. Key observations include: (1) without the deterministic component, both translation and rotation performance dropped significantly; (2) determining the IB weight γ presents a trade-off between the accuracy of translation and rotation prediction; (3) a larger sample size reduces both the uncertainty and prediction errors; and (4) 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.

5.6.1 Effect of the Deterministic Component

We conducted stochastic-only ablation experiments to examine the effects of the deterministic components in Eqs. (9) and (11) by removing the deterministic nodes in Fig. 1b. 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 summarized in Table 5. Without the deterministic component, both translation and rotation performance dropped significantly, which supports the effectiveness of the proposed deterministic component.

Remark. For the 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 tran-

Table 5 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

sitioned state context for the transition model at the next time step. More details of the two implementations are given 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 the 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.

5.6.2 Effect of the IB Weight

We examined the effect of the IB weight, i.e. γ in Eqs. (1) and (4). As shown in Table 6, Although $\gamma = 0.1$ presents a good choice for training on the EuRoC dataset, we observed that the translation and rotation results did not change consistently with different IB weights on the KITTI dataset. While the translation accuracy degrades under a larger γ, the rotation result improves instead. This finding indicates that the determination of the IB weight actually presents a trade-off between the accuracy of translation and rotation predictions and should be taken into account in different scenarios according to the requirements of specific tasks.

5.6.3 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 be 0.01 to avoid a trivial solution, which sets an empirical lower bound of the uncertainty. Table 7 shows that a larger sample size reduces both the uncertainty and prediction errors. An interesting observation from our results is that though more training samples still benefit the prediction performance, the averaged 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 6 Results of varying IB weights γ for InfoVIO

γ	KITTI		EuRoC	
	$t(m)$	$r(^o)$	$t(m)$	$r(^o)$
0.0	0.0639	0.0482	0.0278	0.0814
0.01	0.0559	0.0449	0.0277	0.0794
0.05	0.0570	0.0424	0.0283	0.0785
0.1	0.0580	0.0416	0.0276	0.0744
0.5	0.0612	0.0411	0.0335	0.0765
1.0	0.0648	0.0375	0.0873	0.0948

Table 7 Results of varying sample sizes on KITTI

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

r_n: the ratio of training samples. $\bar{\sigma}^2$: the averaged variance of the latent representation

Table 8 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

5.6.4 Effect of Extra Sensors

Motivated by Theorem 2 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 8, which implies 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 the 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.

5.7 What Does the Intrinsic Uncertainty Mean?

We next used the averaged 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, e.g., larger turning angles and smaller forward distances lead to higher uncertainty. Our analysis suggests a practical data collection

Table 9 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

The best results for each method are shown in bold

Table 10 Results of the proposed intrinsic uncertainties under different data degradation settings on KITTI and EuRoC

	Image	IMU	$\bar{\sigma}^2$ (KITTI)	$\bar{\sigma}^2$ (EuRoC)
Clean	✓	✓	0.0101	0.0103
Noisy	$\mathcal{N}$	✓	0.0102	0.0103
Noisy	✓	$\mathcal{N}$	0.0104	0.0119
Noisy	$\mathcal{N}$	$\mathcal{N}$	0.0104	0.0119
Missing	$\mathcal{M}$	✓	0.0101	0.0103
Missing	✓	$\mathcal{M}$	0.0106	0.0119
Missing	$\mathcal{M}$	$\mathcal{M}$	0.0107	0.0119

$\bar{\sigma}^2$: the averaged variance of the latent representation. ✓, $\mathcal{N}$, and $\mathcal{M}$ denote clean, noisy, and missing data, respectively

guideline, i.e., augmenting the uncertain parts of the pose distribution.

5.7.1 Uncertainty on KITTI and EuRoC

We show the uncertainty results of InfoVIO on KITTI and EuRoC in Figs. 4 and 5, respectively. Since the translations along x and y axes and the rotations around x and z axes are relatively small in the KITTI dataset, their uncertainties do not exhibit a clear pattern. While for the translation along the forward axis-z and the rotation around the upward axis-y (turning left/right), a clear negative and a clear positive relationship are 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 in achieving accurate predictions. 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 relative difficulties in accurately predicting MAV translations since EuRoC has a much smaller translation scale than KITTI.

Remark. There is also a line of work that attempts to combine learning based methods with geometric 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 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 geometric methods.

5.7.2 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 also affected the intrinsic uncertainty, as shown in Figs. 4 and 5. 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 9, 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. Therefore, our findings supports that InfoVO is able to capture this kind of epistemic uncertainty, which is caused by the model design rather than input data. Based on this observation, we propose a simple refinement strategy that eliminates results from the most uncertain position (pos-0) and averagely ensembles the results from the rest positions. We report the refined evaluation results for all models in our main results and ablation studies.

5.7.3 Failure-Awareness

We show that our intrinsic uncertainty measure is failure-aware, which is crucial for a robust odometry system. 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 the test dataset to create noisy data. To generate missing data, we replace the observations with the Gaussian noise.

In Fig. 6, we report the visualization results of uncertainties versus different translations and rotations on KITTI by

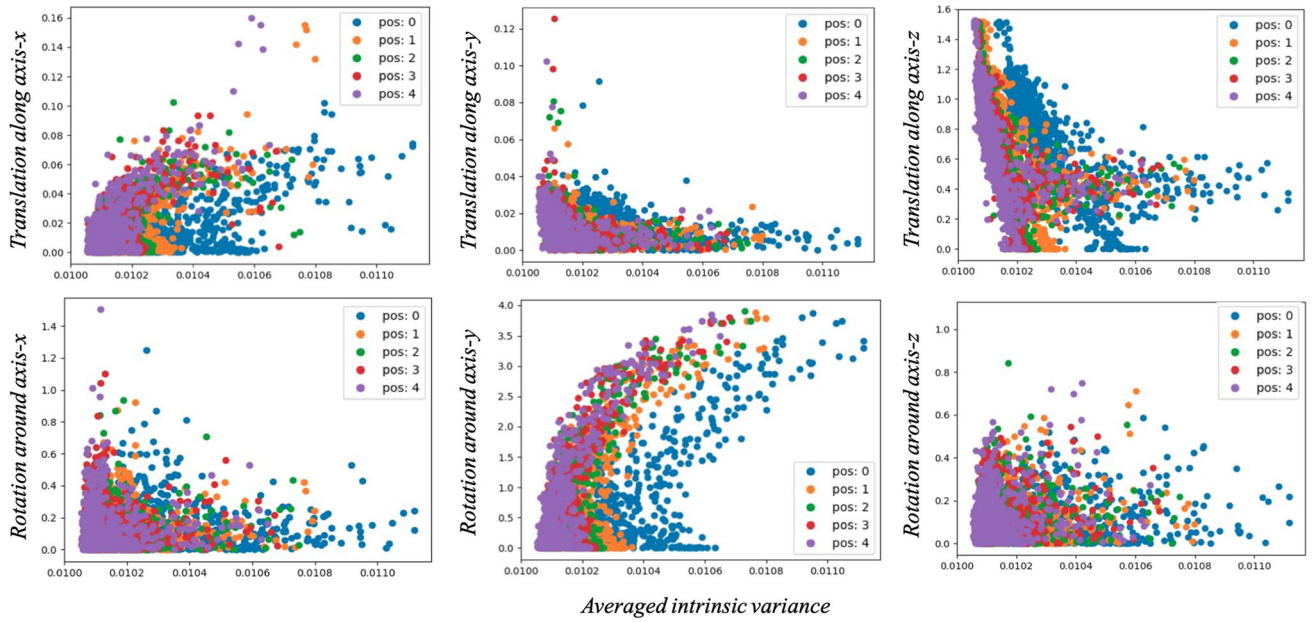


Fig. 4 Uncertainty results of InfoVIO on KITTI. The top and bottom rows represent translation and rotation results. The first, second, and third columns represent x, y, and z, respectively. x, y, z are with respect to the coordinate system in KITTI. pos-i means the result is evaluated at the i-th position in a clip

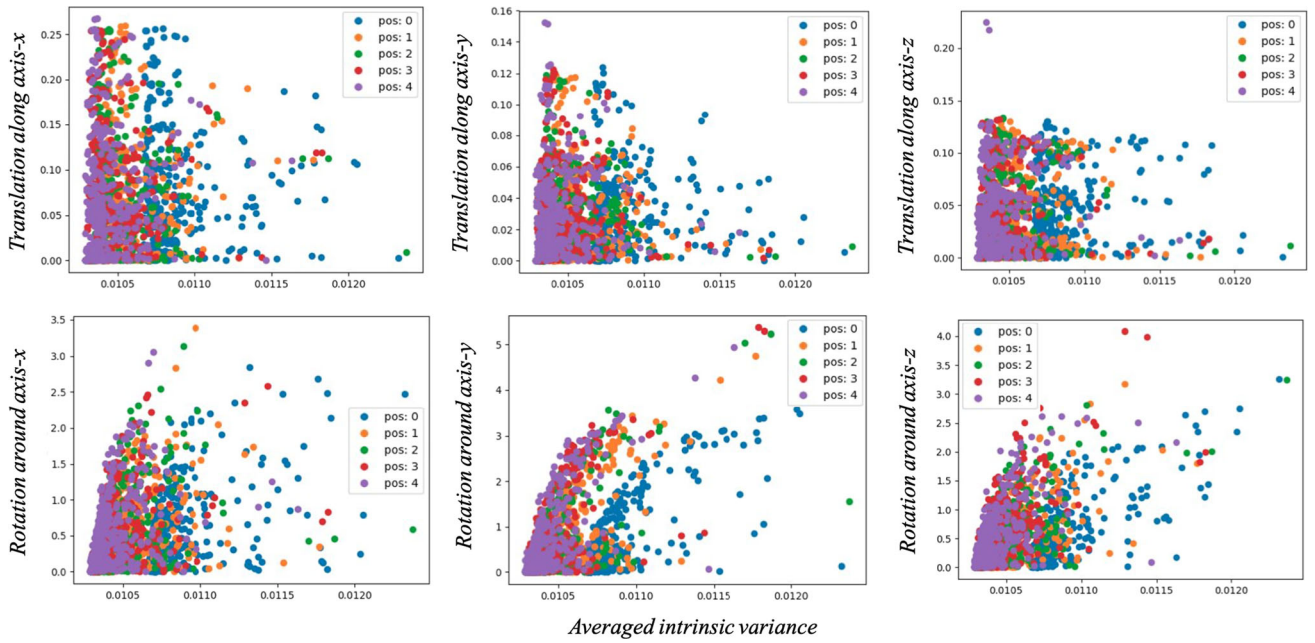


Fig. 5 Uncertainty results of InfoVIO on EuRoC. The arrangement and notation are kept the same as Fig. 4

applying data corruption to both images and IMU. The results of single sensor corruption under the noisy and missing data settings are also provided in Figs. 7 and 8, respectively. The visualization results on EuRoC is provided in the Supplementary Material. We summarize the intrinsic variances under different data degradation settings in Table 10. Our model 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 the visual information. Also, data degradation on IMU records leads to higher uncertainty in EuRoC than in KITTI. We suspect the reason is that the synchronization between the ground-truth poses and IMU records are less accurate in

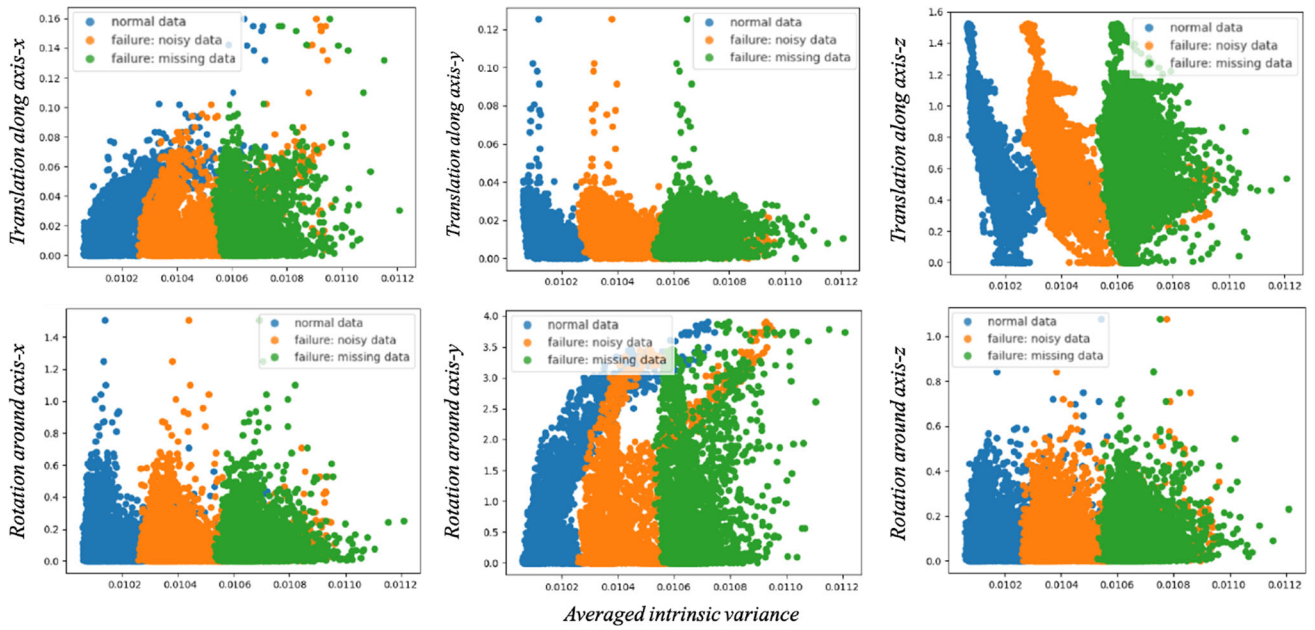


Fig. 6 Uncertainty results of InfoVIO on both noisy and missing data of the KITTI dataset. The arrangement and notation are kept the same as Fig. 4. Blue, orange, and green circles denote results from normal data, noisy data, and missing data, respectively. Both images and IMU records were degraded (Color figure online)

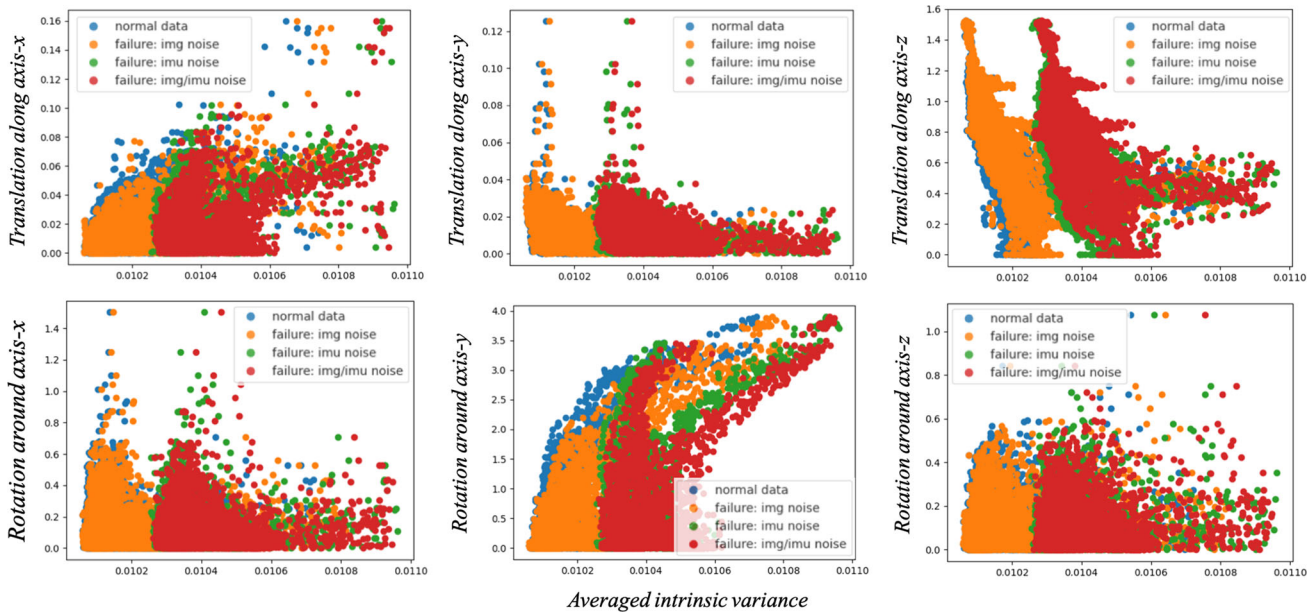


Fig. 7 Uncertainty results of InfoVIO on noisy data of the KITTI dataset. The arrangement and notation are kept the same as Fig. 4. Blue, orange, green, and red circles denote results from normal data and degraded data with images, IMU, and both images and IMU being noisy, respectively (Color figure online)

KITTI than in EuRoC, leading to noisy IMU data for training. At last, the model trained on EuRoC exhibits the same performance on the noisy and the missing data, which implies that EuRoC dataset may be more prone to noises. These observations support that the proposed intrinsic uncertainty measure provides a practical tool for failure diagnoses, such as noises, sensor malfunctions, and even mis-synchronization between sensors.

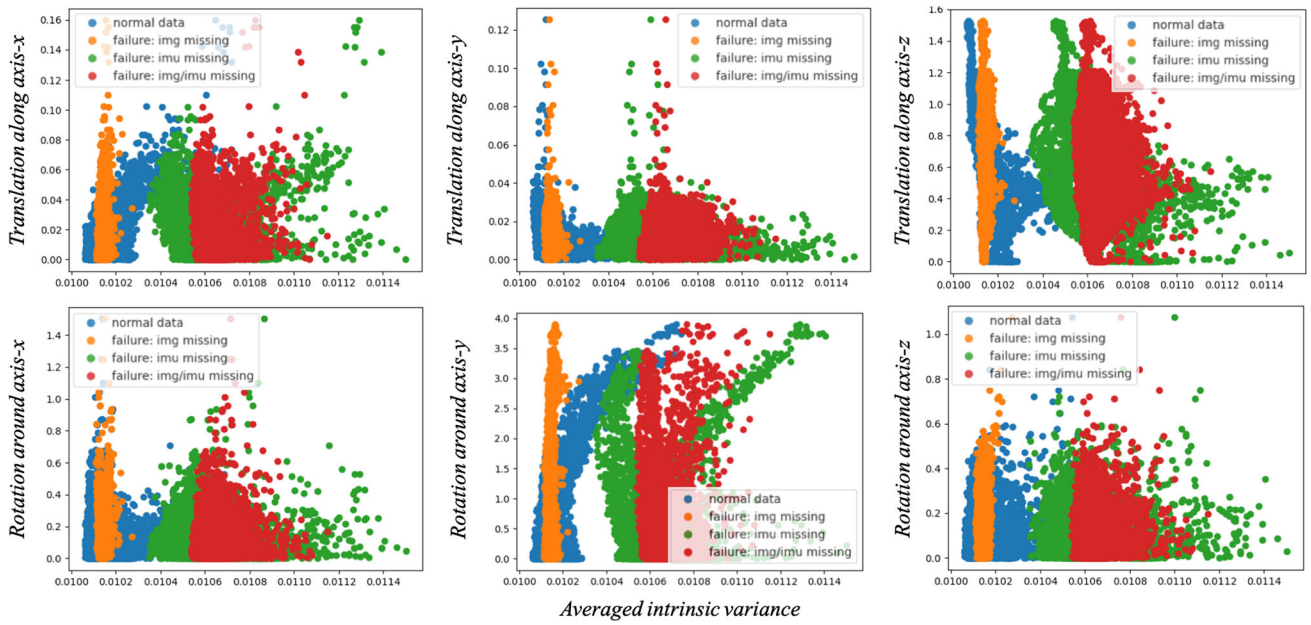


Fig. 8 Uncertainty results of InfoVIO on missing data of the KITTI dataset. The arrangement and notation are kept the same as Fig. 4. Blue, orange, green, and red circles denote results from normal data and degraded data with images, IMU, and both images and IMU missing, respectively

6 Conclusion and Future Research

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.

The proposed method falls into end-to-end supervised learning methods. Obtaining the required ground-truth pose labels can be challenging for large-scale data collection and training. Two recent research trends provide promising solutions to mitigate this problem, i.e. embodied methods that utilize simulated environments and unsupervised learning methods that leveraged the geometric constraints and trained the model jointly with other auxiliary tasks like depth prediction. The difficulty in bringing embodied methods into current state-of-the-art frameworks is the domain gap between simulation and the real world, where proper domain adaptation techniques are desired. Integrating unsupervised and supervised methods can also be challenging, which requires more dedicated training strategies and model design. It is worth noting that our proposed IB method improves on the representation level and can also be applied in these fields to obtain better latent representations. We foresee further developments by incorporating novel techniques into our IB framework.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-022-01659-9.

Acknowledgements This work is supported by ARC FL-170100117, DP-180103424, IC-190100031, and LE-200100049.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

References

Alexander, A. A., Fischer, I., Dillon, J. V., & Murphy, K. (2017). Deep variational information bottleneck. In *ICLR 2017: International conference on learning representations 2017*.

Bian, J., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M. M., & Reid, I. (2019). Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *NeurIPS 2019: 33rd conference on neural information processing systems* (pp. 35–45).

Buesing, L., Weber, T., Racaniere, S., Eslami, S. M., Rezende, D., Reichert, D. P., Viola, F., Besse, F., Gregor, K., Hassabis, D., & Wierstra, D. (2018). Learning and querying fast generative models for reinforcement learning. arXiv preprint arXiv:1802.03006.

Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., Achtelik, M. W., & Siegwart, R. (2016). The Euroc micro aerial vehicle datasets. *The International Journal of Robotics Research, 35*(10), 1157–1163.

Cabon, Y., Murray, N., & Humenberger, M. (2020). Virtual kitti 2. arXiv preprint arXiv:2001.10773.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., & Zecchina, R. (2017). Entropy-sgd: Biasing gradient descent into wide valleys. *International Conference on Learning Representations, 2019*(12), 124018.

Chen, C., Rosa, S., Miao, Y., Lu, C. X., Wu, W., Markham, A., & Trigoni, N. (2019). Selective sensor fusion for neural visual-inertial odometry. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 10542–10551).

Chen, C., Wang, B., Lu, C. X., Trigoni, N., & Markham, A. (2020). A survey on deep learning for localization and mapping: Towards the age of spatial machine intelligence. arXiv preprint arXiv:2006.12567.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). 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).

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C. & Bengio, Y. (2015). A recurrent latent variable model for sequential data. In *NIPS'15 proceedings of the 28th international conference on neural information processing systems* (Vol. 2, pp. 2980–2988).

Clark, R., Wang, S., Wen, H., Markham, A., & Trigoni, N. (2017). Vinet: Visual inertial odometry as a sequence to sequence learning problem. In *31st AAAI conference on artificial intelligence* (pp. 3995–4001).

Thomas, M. T., & Joy, A. T. (1991). *Elements of information theory*. Wiley-Interscience.

Dai, B., Zhu, C., & Wipf, D. (2018). Compressing neural networks using the variational information bottelneck. In *ICML 2018: 35th international conference on machine learning* (pp. 1135–1144).

DDosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., & Brox, T. (2015). Flownet: Learning optical flow with convolutional networks. In *2015 IEEE international conference on computer vision (ICCV)* (pp. 2758–2766).

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

Fuentes-Pacheco, J., Ruiz-Ascencio, J., & Rendón-Mancha, J. M. (2015). Visual simultaneous localization and mapping: A survey. *Artificial Intelligence Review, 43*(1), 55–81.

Gal, Y., & Ghahramani, Z. (2016). 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* (Vol. 48, pp. 1050–1059).

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

Gilitschenski, I., Sahoo, R., Schwarting, W., Amini, A., Karaman, S., & Rus, D. (2019). Deep orientation uncertainty learning based on a Bingham loss. In *International conference on learning representations*.

Goyal, A., Islam, R., Strouse, D., Ahmed, Z., Botvinick, M., Larochelle, H., Bengio, Y., & Levine, S. (2019). Infobot: Transfer and exploration via the information bottleneck. In *ICLR 2019: 7th international conference on learning representations*.

Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., & Davidson, J. (2019). Learning latent dynamics for planning from pixels. In *ICML 2019: 36th international conference on machine learning* (pp. 2555–2565).

Hafner, D., Lillicrap, T., Ba, J., & Norouzi, M. (2020). Dream to control: learning behaviors by latent imagination. In *ICLR 2020: 8th international conference on learning representations*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hu, J.-S., & Chen, M.-Y. (2014). 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).

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., & Brox, T. (2017). 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).

Kendall, A., & Gal, Y. (2017). 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).

Kendall, A., Grimes, M., & Cipolla, R. (2015). Posenet: A convolutional network for real-time 6-dof camera relocalization. In *2015 IEEE international conference on computer vision (ICCV)* (pp. 2938–2946).

Khatri, C. G., & Mardia, K. V. (1977). The von Fises–Fisher matrix distribution in orientation statistics. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 95–106.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In *ICLR 2014: International conference on learning representations (ICLR) 2014*.

Leutenegger, S., Lynen, S., Bosse, M., Siegwart, R., & Furgale, P. (2015). Keyframe-based visual-inertial odometry using nonlinear optimization. *The International Journal of Robotics Research, 34*(3), 314–334.

Loquercio, A., Segu, M., & Scaramuzza, D. (2020). A general framework for uncertainty estimation in deep learning. *IEEE International Conference on Robotics and Automation (ICRA), 5*(2), 3153–3160.

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

Mourikis, A. I., & Roumeliotis, S. I. (2007). A multi-state constraint kalman filter for vision-aided inertial navigation. In *Proceedings 2007 IEEE international conference on robotics and automation* (pp. 3565–3572). IEEE.

Peretroukhin, V., & Kelly, J. (2017). Dpc-net: Deep pose correction for visual localization. *International Conference on Robotics and Automation, 3*(3), 2424–2431.

Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., & Tucker, G. (2019). On variational bounds of mutual information. In *ICML 2019: 36th international conference on machine learning* (pp. 5171–5180).

Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., & Black, M. J. (2019). Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion seg-

mentation. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 12240–12249).

Shwartz-Ziv, R., & Tishby, N. (2017). Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810.

Steiner, B., DeVito, Z., Chintala, S., Gross, S., Paszke, A., Massa, F., Lerer, A., Chanan, G., Lin, Z., Yang, E., Desmaison, A., Tejani, A., Kopf, A., Bradbury, J., Antiga, L., Raison, M., Gimelshein, N., Chilamkurthy, S., Killeen, T., Fang, L., & Bai, J. (2019). Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS 2019: 33rd conference on neural information processing systems* (pp. 8026–8037).

Sturm, J., Engelhard, N., Endres, F., Burgard, W., & Cremers, D. (2012). A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems* (pp. 573–580).

Taketomi, T., Uchiyama, H., & Ikeda, S. (2017). Visual slam algorithms: A survey from 2010 to 2016. *IPSJ Transactions on Computer Vision and Applications, 9*(1), 16.

Teed, Z., & Deng, J. (2020). Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision* (pp. 402–419). Springer.

Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *2015 IEEE information theory workshop (ITW)* (pp. 1–5).

Tishby, N., Pereira, F. C. N., & Bialek, W. (2000). The information bottleneck method. In *Proc. 37th annual Allerton conference on communications, control and computing, 1999* (pp. 368–377).

Vera, M., Piantanida, P., & Vega, L. R. (2018). The role of the information bottleneck in representation learning. In *2018 IEEE international symposium on information theory (ISIT)* (pp. 1580–1584).

Wang, S., Clark, R., Wen, H., & Trigoni, N. (2017). 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).

Wang, S., Clark, R., Wen, H., & Trigoni, N. (2018). End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks. *The International Journal of Robotics Research, 37*, 513–542.

Xu, A., & Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms. In *31st annual conference on neural information processing systems, NIPS 2017* (pp. 2524–2533).

Xu, H., Zhang, J., Cai, J., Rezatofighi, H., & Tao, D. (2022). Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8121–8130).

Xu, Y., Zhang, Q., Zhang, J., & Tao, D. (2021). Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems, 34*, 28522–28535.

Xue, F., Wang, X., Li, S., Wang, Q., Wang, J., & Zha, H. (2019). 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).

Yang, N., von Stumberg, L., Wang, R., & Cremers, D. (2020). D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1281–1292).

Yin, Z., & Shi, J. (2018). 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).

Zhan, H., Weerasekera, C. S., Bian, J. W., & Reid, I. (2020). Visual odometry revisited: What should be learnt? In *2020 IEEE international conference on robotics and automation (ICRA)* (pp. 4203–4210). IEEE.

Zhang, J., & Tao, D. (2020). Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. *IEEE Internet of Things Journal, 8*(10), 7789–7817.

Zhang, J., Liu, T., & Tao, D. (2021). An optimal transport analysis on generalization in deep learning. *In IEEE Transactions on Neural Networks and Learning Systems* (pp. 1–12).

Zhang, Q., Xu, Y., Zhang, J., & Tao, D. (2022a). Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. arXiv preprint arXiv:2202.10108.

Zhang, S., Zhang, J., & Tao, D. (2022b). Towards scale-aware, robust, and generalizable unsupervised monocular depth estimation by integrating IMU motion dynamics. arXiv preprint arXiv:2207.04680.

Zhang, S., Zhang, J., & Tao, D. (2022c). Towards scale consistent monocular visual odometry by learning from the virtual world. arXiv preprint arXiv:2203.05712.

Zhou, T., Brown, M., Snavely, N., & Lowe, D. G. (2017). Unsupervised learning of depth and ego-motion from video. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 6612–6619).