Quantifying Epistemic Uncertainty in Absolute Pose Regression

Fereidoon Zangeneh[1,2(✉)], Amit Dekel[2], Alessandro Pieropan[2], and Patric Jensfelt[1]

[1] KTH Royal Institute of Technology, Stockholm, Sweden
{fzk,patric}@kth.se
[2] Univrses AB, Stockholm, Sweden
{fereidoon.zangeneh,amit.dekel,alessandro.pieropan}@univrses.com

Abstract. Visual relocalization is the task of estimating the camera pose given an image it views. Absolute pose regression offers a solution to this task by training a neural network, directly regressing the camera pose from image features. While an attractive solution in terms of memory and compute efficiency, absolute pose regression's predictions are inaccurate and unreliable outside the training domain. In this work, we propose a novel method for quantifying the epistemic uncertainty of an absolute pose regression model by estimating the likelihood of observations within a variational framework. Beyond providing a measure of confidence in predictions, our approach offers a unified model that also handles observation ambiguities, probabilistically localizing the camera in the presence of repetitive structures. Our method outperforms existing approaches in capturing the relation between uncertainty and prediction error.

Keywords: Camera Relocalization · Uncertainty Estimation · VAEs

1 Introduction

Visual localization is the task of estimating the pose of a camera from the image that it captures in the environment. It is a technique used for indoor navigation of robots, augmented reality devices, and autonomous driving [5]. A full visual localization pipeline consists of both frame-to-frame tracking of the camera—or relative pose estimation, as well as global relocalization—or absolute pose estimation. The latter refers to estimating the camera pose from a single image in a previously mapped environment, and is the focus of this work. Visual relocalization has been a long-standing topic of research [8,33,34], with works in search of map representations and algorithms that can most efficiently and accurately estimate the camera pose. Traditional solutions range from image databases to sparse 3D point models of the world as their map representations, used for image retrieval or keypoint matching to estimate the pose for a novel query image [1,13]. These solutions face a trade-off between efficiency and accuracy. A more recent paradigm for visual relocalization relies on using end-to-end trainable

J. Petersen and V. A. Dahl (Eds.): SCIA 2025, LNCS 15726, pp. 180–195, 2025.
https://doi.org/10.1007/978-3-031-95918-9_13

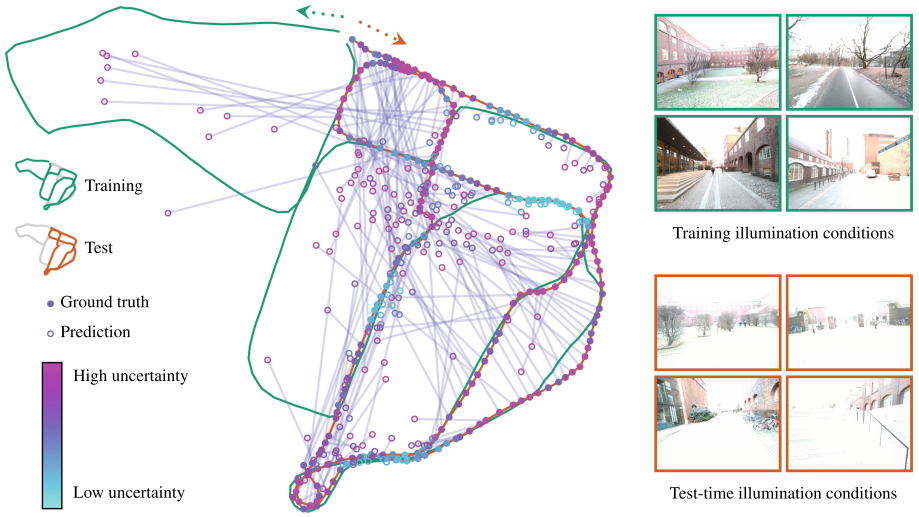

Fig. 1. Querying any absolute pose regression network on a trajectory and data domain different from training data results in high prediction errors (depicted by long lines connecting predictions and ground truth). Our proposed epistemic uncertainty quantification approach estimates the likelihood of test samples belonging to the training distribution (visualized by a color map), offering a guide for trusting the predictions. We see that the color of the predictions and their ground truths, which encodes the uncertainty, is highly correlated with the prediction error (the length of the lines).

pipelines to directly regress the camera pose from the image by a neural network [3,17,35,36]. These end-to-end methods store a map representation in the weights of a neural network, promising higher memory and compute efficiency, as well as better robustness compared to the traditional methods [39].

Although end-to-end absolute pose regression methods are appealing for their efficiency, they lag behind traditional geometric approaches in terms of accuracy, as evidenced by their performance on visual relocalization benchmarks [30,37]. While the initial motivation for using neural networks in this task was their potential to learn a spatial understanding of the scene, in practice, these models often perform similarly to pose approximations based on image retrieval [32], limiting their generalization beyond the training data. This limitation is further exacerbated by the absence of geometric constraints or verification steps in their prediction process—every input yields a prediction, which can be highly inaccurate. This contrasts with traditional methods, which can identify when there is insufficient evidence to make a reliable prediction. Despite these limitations, the fast inference speed and small memory footprint of pose regression networks has driven significant research efforts aimed at addressing its shortcomings by improving its generalization [25] and accuracy [6,35]. A remaining challenge in bridging the gap between the current state of absolute pose regression and real-world applications is to accompany network predictions with a reliable measure of confidence. Figure 1 illustrates a real-world scenario where a network is tested

on a trajectory significantly different from the training/mapped region. The scenario also suffers from strong lighting variations between training and testing, further straining the visual relocalization system. In such cases, the network needs to *know when it does not know*.

Existing research on uncertainty estimation in absolute pose regression has followed two distinct tracks: one focuses on handling ambiguous observations by modeling aleatoric uncertainty [9,24,42], while the other—including this paper—aims to quantify epistemic uncertainty to provide confidence in predictions [12,14]. However, the methodologies used in these two tracks differ fundamentally leading to different frameworks. In this work, we build on a solution based on variational autoencoders (VAE) [43], originally proposed to address the challenge of handling ambiguous observations from repetitive structures. We show that VAEs can also be used for quantifying epistemic uncertainty, resulting in a unified framework that can handle both natures of uncertainty. We propose a novel method for quantifying epistemic uncertainty in absolute pose regression to estimate the reliability of the predictions, as shown in Fig. 1.

In summary: (1) We propose a unified framework for handling epistemic as well as aleatoric uncertainty in absolute pose regression. (2) We derive a formulation for quantification of epistemic uncertainty in absolute pose regression within a variational framework. (3) We show how our epistemic uncertainty quantification can estimate the reliability of network predictions. (4) We perform a thorough evaluation to show that our method outperforms the existing epistemic uncertainty quantification methods.

2 Related Work

2.1 Visual Relocalization

Traditional Methods. The well-established solutions to visual relocalization fall into two paradigms: image-based and structure-based methods. Image-based methods store a database of images and their camera poses as a representation of the map. Given a query image, they rely on global image descriptors [1] to retrieve the most similar image(s) in the database, hence approximate its pose [38]. The map representation in structure-based methods is a sparse 3D point model of the scene. For localizing a query image, 2D-3D matching is performed between the image and the sparse model, enabling accurate estimation of the camera pose [21,31]. These two paradigms have distinct strengths and properties on the accuracy-efficiency trade-off, such that they are typically combined in a hierarchical framework for large-scale relocalization [30]. Both paradigms, however, exhibit detectable signs of failure when queried with an image that deviates significantly from the map; image-based retrieval computes too little similarity between query and the map, and structure-based matching finds too small inlier sets. So each estimation can be accompanied with a measure of reliability.

Regression-Based Methods. A recent trend in visual relocalization is using end-to-end learning-based pipelines. These methods encode a map representation

in the weights of a neural network that directly regresses a geometric quantity from an image. This makes them appealing alternatives to traditional methods for their space and computational efficiency in large scenes, along with data-driven handling of lighting changes, camera blur and texture-less surfaces [39]. These methods come in two main variants: scene coordinate regression and absolute pose regression. In scene coordinate regression, the neural network regresses the scene's 3D point coordinates from image patches, which are then used in a robust estimation process to compute the camera pose [36]. In absolute pose regression, the neural network directly predicts the camera pose, eliminating the need for the robust estimation step [17]. Both variants are appealing in their own right. Scene coordinate regression, due to its more local nature of predictions, shows higher generalization capabilities than absolute pose regression, resulting in more accurate predictions [3]. This is particularly beneficial in presence of occlusions, where scene coordinate regression can leverage image patches that provide consistent information to produce an accurate final pose, while the global pose predicted by an absolute pose regression network may be adversely affected. Scene coordinate regression, however, originally proposed for relocalization with RGB-D images, requires additional, more complex steps for end-to-end training with RGB images [2,3]. In contrast, absolute pose regression comes with a simpler training procedure, and despite its shortcomings [32] remains an attractive alternative. This has promoted research efforts to build upon the original work [17], including the development of more effective geometric and photometric loss functions [4,7,15], network architectures [23,39,41], extension to multiple scenes [35], and improved generalization through data synthesis during training [25–27]. Absolute pose regression in its basic form produces a point prediction for any given input through a forward pass of the network, without providing any confidence in the prediction. A complementary line of research centers on incorporating uncertainty estimation into pose regression networks, enhancing their reliability for real-world applications. Our work aligns with this direction.

2.2 Uncertainty Estimation in Absolute Pose Regression

Uncertainty in neural network predictions arises in two distinct forms: aleatoric uncertainty and epistemic uncertainty [16]. Aleatoric uncertainty arises from inherent noise and ambiguity in the data, which cannot be reduced even with an infinite amount of training data. An example of this in camera relocalization is repetitive structures that appear visually similar but correspond to different camera poses. Epistemic uncertainty, on the other hand, reflects the model's lack of knowledge about a data point—specifically, how well that point fits within the distribution learned during training.

Modeling Aleatoric Uncertainty. Aleatoric uncertainty is typically classified into two types based on its dependence on the input data: heteroscedastic uncertainty, which varies with the input; and homoscedastic uncertainty, which remains constant regardless of the input such as sensor noise. Kendall and Cipola [15] used the latter paradigm to learn the optimal weighting of translation and

orientation error terms during training of each scene. On the other hand, heteroscedastic uncertainty is modeled to account for ambiguities in observations. A common method for modeling heteroscedastic aleatoric uncertainty is by predicting a unimodal parametric distribution [16]. Moreau et al. [24] applied this technique to absolute pose regression in outdoor environments, integrating it in a tracking loop. To address multimodal posterior distributions caused by repetitive structures, Deng et al. [9] proposed predicting a mixture model instead of a unimodal distribution. In contrast, Zangeneh et al. [42] adopted a non-parametric approach, representing the predicted distribution through samples rather than a mixture model. This approach was later reformulated with a conditional VAE in [43], where a VAE is trained to reconstruct camera pose samples conditioned on image observations. As a result of this, the latent space captures the space of pose ambiguities given an observation.

Quantifying Epistemic Uncertainty. A number of works have explored the quantification of epistemic uncertainty in absolute pose regression. The pioneering work of Kendall and Cipolla [14] follows Bayesian deep learning principles and incorporates dropout layers in the network to approximate variational inference for the posterior distribution of model weights [10]. At test time, Monte Carlo dropout is applied, performing stochastic forward passes over a query image, with the variance of predictions providing a measure of epistemic uncertainty. The use of dropout to generate multiple predictions was later adopted by [12], with the distinction that instead of applying dropout within a Bayesian neural network, it was applied directly to the input image pixels. This approach served a dual purpose of quantifying uncertainty along with enhancing robustness against moving objects, aim at autonomous driving scenarios. More recently, Liu et al. [20] proposed a functional method for uncertainty quantification, which directly compares the image features and the predicted pose of a query to those of the training samples. The presence of a training sample similar to the query, in both pose and appearance, is interpreted as low epistemic uncertainty for the model, with the intuition that the model must be well-informed on the query.

While conditional VAEs have been previously explored to explicitly model the aleatoric uncertainty of pose regression models [43], they can in theory also be used for likelihood estimation of test samples [18], and thus implicitly capture model's epistemic uncertainty about the prediction. Specifically, the structured latent space and the agreement between the encoder and decoder that is learned for a training set can be used to detect samples that are unlikely under the training distribution. In this work, we explore this capability of conditional VAEs for epistemic uncertainty quantification in absolute pose regression, assessing its viability as an alternative to existing methods while providing a unified framework to also capture aleatoric uncertainty.

3 Method

An ideal visual relocalization solution predicts the true camera pose $y \in \mathrm{SE}(3)$ for any given image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}$ sampled in the scene, where $(\boldsymbol{x}, y) \sim p_{\text{true}}$.

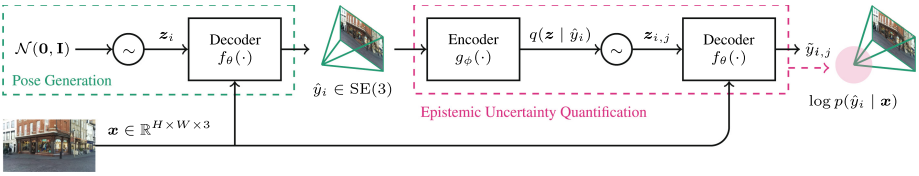

Fig. 2. Our pipeline models a scene in a conditional VAE setup. Given a test-time image observation $\boldsymbol{x}$, the decoder is used to sample poses $\hat{y}$ from the posterior distribution of camera poses $p(y \mid \boldsymbol{x})$. Reconstruction of $\hat{y}$ through the VAE pipeline then gives an estimate on $\log p(\hat{y} \mid \boldsymbol{x})$, that is the likelihood of the test sample under the training distribution. This reflects model's epistemic uncertainty about the observation $\boldsymbol{x}$.

In practice, the real-world distribution p_{true} can only be modeled by a finite set of training samples that form an empirical distribution p_{train}. Successful modeling of p_{train} implies generalization to any other unseen empirical distribution p_{test} from the underlying p_{true}. However, there is in practice always a model performance gap between queries from p_{train} and p_{test}. We are interested in a visual relocalization model that once trained, can measure how well a test sample $(\boldsymbol{x}, y) \sim p_{\text{test}}$ adheres to the modeled distribution p_{train}, hence estimate how reliable its predictions are. To this end, we frame the visual relocalization task as learning the distribution $p_{\text{train}}(y \mid \boldsymbol{x})$. Sampling from this conditional distribution performs the task itself [43], while estimating its likelihood for a given sample $(\boldsymbol{x}, y)$ quantifies the degree of its conformity to the modeled distribution p_{train}. We discuss the training procedure to model and sample from $p_{\text{train}}(y \mid \boldsymbol{x})$ in Sect. 3.1. We then lay down our proposed approach to estimate the likelihood of samples and quantify model's epistemic uncertainty in Sect. 3.2. We hereafter refer to $p_{\text{train}}(y \mid \boldsymbol{x})$ as $p(y \mid \boldsymbol{x})$.

3.1 Learning a Generative Model

We intend to train a neural network $f_\theta(\cdot)$ that conditioned on a given image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}$ transforms samples $\boldsymbol{z} \in \mathbb{R}^d$ from a noise distribution $p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ to the posterior distribution over camera poses $y \in \mathrm{SE}(3) \sim p(y \mid \boldsymbol{x})$. A shown in [43], such a generative network can be trained as the decoder in a conditional VAE pipeline that reconstructs camera poses given images in the scene. We briefly outline this setup and training procedure below, while referring the reader to [43] for its design principles.

Setup and Optimization. The conditional VAE consists of an encoder $g_\phi(\cdot)$ and a conditional decoder network $f_\theta(\cdot)$ that are optimized to, together, reconstruct the camera pose y_i for a given image $\boldsymbol{x}_i$ from the training set $(\boldsymbol{x}_i, y_i) \in \mathcal{D}_{\text{train}}$. The encoder maps each training pose y_i to the mean and covariance of a Gaussian posterior $q(\boldsymbol{z} \mid y_i)$ that is intended to resemble the true latent posterior distribution $p(\boldsymbol{z} \mid y_i)$. The decoder, conditioned on the corresponding image $\boldsymbol{x}_i$, then maps samples drawn from this inferred posterior $\boldsymbol{z}_j \sim q(\boldsymbol{z} \mid y_i)$ to reconstructions $\hat{y}_{i,j} \in \mathrm{SE}(3)$ of the original pose y_i.

The optimization objective of the pipeline is the evidence lower bound (ELBO) [19]. From the likelihood of training samples it is possible to derive that

$$\begin{aligned} &\log p(y|\boldsymbol{x}) - \overbrace{D_{\mathrm{KL}}\big(q(\boldsymbol{z} \mid y) \parallel p(\boldsymbol{z} \mid y)\big)}^{\geq 0} \\ &= \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{z}|y)} \log p_\theta(y \mid \boldsymbol{z}, \boldsymbol{x}) - D_{\mathrm{KL}}\big(q_\phi(\boldsymbol{z} \mid y) \parallel p(\boldsymbol{z})\big)}_{\mathrm{ELBO}}, \end{aligned} \tag{1}$$

which indicates that maximizing ELBO with optimization variables ϕ and θ over $\mathcal{D}_{\text{train}}$ optimizes the networks $g_\phi(\cdot)$ and $f_\theta(\cdot)$ that model $p(y \mid \boldsymbol{x})$. In the above expression the subscripts ϕ, θ denote the relation between distributions and weights of the networks they are materialized in. We compute the expected value of the reconstruction likelihood with Monte Carlo samples from $q_\phi(\boldsymbol{z} \mid y)$, following the Gaussian model

$$\log p_\theta\big(y \mid \boldsymbol{z}, \boldsymbol{x}\big) = -1/2\big(6 \log 2\pi + \log \det(\boldsymbol{\Sigma}) + \boldsymbol{\xi}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\xi}\big), \tag{2}$$

where a homoscedastic 6×6 covariance matrix $\boldsymbol{\Sigma}$ is shared for training all samples and optimized alongside network weights θ and ϕ. We take the reconstruction error to be a local correction on the prediction such that $y = \hat{y} \exp(\boldsymbol{\xi}^\wedge)$, hence define the error vector as $\boldsymbol{\xi} = \log(\hat{y}^{-1} y)^\vee \in \mathbb{R}^6$ in the Lie algebra $\mathfrak{se}(3)$[1]. Defining the error in the tangent space and learning a shared $\boldsymbol{\Sigma}$ enables automatic adjustment of loss weights across the different degrees of freedom of the camera pose throughout training. This automatic adjustment is inspired by [15] and eliminates the need for the suboptimal manual tuning performed per dataset [9,42,43]. However, while [15] explored learning only two loss weight parameters between translation and rotation error components, we extend this approach to all six degrees of freedom by modeling the full covariance matrix.

Sample Generation. We can easily sample from $p(y \mid \boldsymbol{x})$ by conditioning the decoder on the image $\boldsymbol{x}$ and passing samples from the prior through the decoder to get $\mathcal{Y} = \{\hat{y}_i = f_\theta(\boldsymbol{z}_i, \boldsymbol{x}) \mid \boldsymbol{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})\}$. This is illustrated in the left dotted box in Fig. 2. As discussed in [43], the decoder learns the space of aleatoric pose ambiguities associated with each image, and for ambiguous images, it appropriately splits the latent space such that different latent regions are mapped to distinct camera poses for an image.

3.2 Likelihood Estimation

We propose to quantify the epistemic uncertainty of a trained model about a test sample by computing its marginal likelihood, which we can estimate by importance sampling:

[1] $\exp : \mathfrak{se}(3) \mapsto \mathrm{SE}(3)$ is the exponential map from Lie algebra to Lie group and log is its inverse. The operator $\wedge$ turns $\boldsymbol{\xi}$ into a member of the Lie algebra $\mathfrak{se}(3)$ and $\vee$ is its inverse.

$$\begin{aligned}\log p(y \mid \boldsymbol{x}) &= \log \int p_\theta(y \mid \boldsymbol{z}, \boldsymbol{x}) \overbrace{p(\boldsymbol{z} \mid \boldsymbol{x})}^{=p(\boldsymbol{z})} d\boldsymbol{z} \\ &= \log \int q_\phi(\boldsymbol{z} \mid y) \frac{p_\theta(y \mid \boldsymbol{z}, \boldsymbol{x}) p(\boldsymbol{z})}{q_\phi(\boldsymbol{z} \mid y)} d\boldsymbol{z} \\ &= \log \mathbb{E}_{q_\phi(\boldsymbol{z}|y)} \frac{p_\theta(y \mid \boldsymbol{z}, \boldsymbol{x}) p(\boldsymbol{z})}{q_\phi(\boldsymbol{z} \mid y)} \\ &\approx \log \frac{1}{M} \sum_{\boldsymbol{z}_j} \frac{p_\theta(y \mid \boldsymbol{z}_j, \boldsymbol{x}) p(\boldsymbol{z}_j)}{q_\phi(\boldsymbol{z}_j \mid y)} \quad \substack{\boldsymbol{z}_j \sim q_\phi(\boldsymbol{z}|y) \\ j=1:M}\end{aligned} \tag{3}$$

The equality $p(\boldsymbol{z} \mid \boldsymbol{x}) = p(\boldsymbol{z})$ stems from the assumption that all observations $\boldsymbol{x}$ share the same prior distribution $p(\boldsymbol{z})$ [40]. At test time, only the image $\boldsymbol{x}$ is available while the true pose y is unknown. Therefore, we propose estimating the expected $\log p(\hat{y}_i|\boldsymbol{x})$ with Monte Carlo generations $\hat{y}_i = f_\theta(\boldsymbol{z}_i, \boldsymbol{x}), \boldsymbol{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This likelihood measure quantifies the agreement between encoder and decoder when conditioned on $\boldsymbol{x}$, which is how a VAE implicitly captures epistemic uncertainty. The two networks are expected to be in agreement only for in-distribution samples, as the latent space is structured based on the training distribution. In other words, we interpret the likelihood $\log p(\hat{y}|\boldsymbol{x})$ as the network's confidence in its prediction for observation $\boldsymbol{x}$. Our proposed pipeline for pose estimation and epistemic uncertainty quantification is outlined in Fig. 2.

4 Implementation Details

We implement the encoder and decoder networks both as multilayer perceptrons with 5 layers of 512 neurons each, and residual connections between their input and third layers. All neurons (except at output) go through LeakyReLU activations. A pretrained ResNet-18 [11] backbone is used to extract 512-dimensional conditioning image feature vectors. The input to the decoder is a concatenation of image feature and latent sample vectors. We opt for a 2-dimensional latent space in all experiments. As the encoder (if expressive enough) is able to learn the surjection of poses to appearances, it is not strictly necessary to condition the encoder on the image features and we obtained similar results for both encoder configurations. We parameterize an SE(3) element at encoder input and decoder output with a 3-vector for translation, and a 6-vector for rotation that can be continuously mapped to a valid rotation matrix by a Gram-Schmidt process [44]. The 2×2 latent posterior covariance matrix is parameterized by its log-variances and an SO(2) rotation. The 6×6 homoscedastic noise covariance matrix is parameterized by its lower triangular matrix from Cholesky decomposition.

We train the networks for 30k iterations with a learning rate of 1e−4 and Adam optimizer with a decoupled weight decay of 1e−2 [22]. We use a batch size of 128 and draw 100 Monte Carlo samples to estimate ELBO's expected reconstruction likelihood term in (1). We found a $0 \rightarrow 1$ warm-up of KL divergence term after 10k iterations to help with better convergence. Following previous works [9,17,43] the images are first resized such that the smaller edge is 256

pixels, then randomly cropped to 244×244 squares during training, and center-cropped to the same size at test time. To improve the conditioning of the decoder output, a linear transformation is computed for each scene, allowing the network to predict values within the $[0, 1]$ range for each translation component, which are then mapped to the true metric scale. To enhance robustness to motion blur and lighting variations, we apply random Gaussian blurring with $\sigma \in [0.05, 5]$ during training, along with brightness and color jittering using PyTorch's [29] ColorJitter function, with brightness, contrast, and saturation set to 0.2, and hue to 0.1.

5 Experiments

We aim to assess the efficacy of our proposed likelihood-based quantification of epistemic uncertainty. Defining a measurable ground truth for epistemic uncertainty is inherently challenging. However, we identify and design an evaluation protocol around a key statistical property that the uncertainties should exhibit. We evaluate our method across a variety of operation conditions, comparing it against existing techniques.

5.1 Evaluation Protocol

Given a trained absolute pose regression model, predictions tend to diverge from the ground truth when test samples deviate from the training distribution. In other words, we expect to observe higher prediction errors for test samples that the model is less informed about and assigns lower likelihood values for. Therefore, the quantified uncertainty should ideally increase with the prediction error. While the exact nature of this relationship remains unknown, we can except one to be a monotonically increasing function of the other. We evaluate this by computing the Spearman's rank correlation coefficient between the estimated negative log-likelihood and prediction error of test samples. This coefficient ranges from -1 to $+1$, where $+1$ is the ideal scenario where negative log-likelihood (epistemic uncertainty) is a perfect monotonically increasing function of prediction error. Through this score, we offer a quantitative counterpart to the qualitative uncertainty-error tables and plots used in [12,20]. As model predictions are made in SE(3) that does not have a natural metric to measure prediction errors, we report the correlation coefficients for translation and rotation error separately.

5.2 Comparison Baselines

We compare our method with existing approaches for epistemic uncertainty estimation in absolute pose regression: Bayesian PoseNet (BPN) [14], RVL [12], and HR-APR [21]. For a fair comparison and to rule out the predictive power of different architectures as a factor, we train all methods with the same network architecture and training setup as our own. Given that RVL's dropout statistics are primarily suited for driving scenarios, which we do not evaluate, we use a

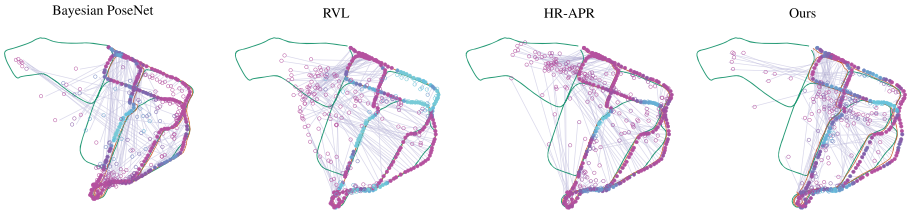

Fig. 3. Testing sequence `kth_day_09` on models trained on `kth_day_06` from [28]. The lines depict the translation error between the predictions (○) and ground truth (●), which are colored from cyan to magenta by the epistemic uncertainty. The length of the lines is expected to correlate with epistemic uncertainty, as is the case in our predictions, that is cyan for short lines (small error) to magenta for long lines (large error). (Color figure online)

uniform distribution for their dropout instead. We obtain the distribution of predicted poses for BPN and RVL by 100 Monte Carlo forward passes. While they propose variance of predictions as the measure of epistemic uncertainty, we found that using negative log-likelihood of the mean prediction benefits both methods in our evaluation metric. So we opt to use that instead to quantify their uncertainties. HR-APR quantifies epistemic uncertainty by a visual dissimilarity measure, which saturates for positionally unlikely samples.

5.3 Datasets

We evaluate our method on the test splits of the real-world datasets Cambridge Landmarks [17], 7-Scenes [36], and ambiguous relocalization sequences of [9,42], covering a variety of outdoor and indoor environments. Additionally, we use two sequences from the Multi-Campus Dataset [28] to evaluate the performance of absolute pose regression and uncertainty quantification on test samples far outside the training trajectory.

Table 1. Median prediction error—translation and rotation separately (lower is better)

Dataset	BPN [14]		RVL [12]		HR-APR [20]		Ours	
	Tra / m	Rot / ○	Tra / m	Rot / ○	Tra / m	Rot / ○	Tra / m	Rot / ○
Ambiguous [9] + Ceiling [42]	0.94	13.13	0.87	28.15	1.36	15.47	**0.21**	**9.06**
7-Scenes [36]	**0.19**	**9.48**	0.23	10.98	0.23	11.05	0.23	11.46
Cambridge Landmarks [17]	**0.67**	11.61	1.23	7.69	1.29	7.38	1.21	**7.17**
Multi-Campus: KTH Day [28]	**22.73**	61.20	29.84	36.02	23.34	43.73	24.28	**32.24**

Table 2. Spearman's rank correlation coefficient for quantified epistemic uncertainties and prediction errors—translation and rotation separately (−1 to +1, higher is better)

Scene	BPN [14]		RVL [12]		HR-APR [20]		Ours	
	Tra	Rot	Tra	Rot	Tra	Rot	Tra	Rot
Ceiling	0.16	**0.68**	0.18	0.59	0.25	0.01	**0.38**	0.21
Blue Chairs	0.02	−0.07	−0.46	0.49	0.28	0.23	**0.64**	**0.65**
Meeting Table	0.08	0.01	−0.12	−0.11	**0.21**	**0.20**	0.16	−0.01
Seminar	0.24	0.14	**0.82**	**0.82**	0.56	0.32	0.76	0.56
Staircase	0.24	0.39	0.47	0.53	0.55	0.55	**0.67**	**0.64**
Staircase Extra	0.10	0.20	0.19	**0.28**	0.24	0.19	**0.26**	0.09
Chess	0.25	−0.17	0.43	0.31	0.16	0.32	**0.43**	**0.34**
Fire	−0.06	0.13	**0.57**	0.60	0.45	**0.62**	0.38	0.53
Heads	0.03	−0.27	0.43	0.31	0.52	0.58	**0.70**	**0.61**
Office	0.25	0.19	**0.50**	0.45	0.48	**0.49**	0.41	0.44
Pumpkin	0.12	−0.07	**0.53**	0.46	0.44	**0.47**	0.50	0.43
Redkitchen	0.01	−0.02	**0.55**	**0.52**	0.50	0.39	0.48	0.51
Stairs	−0.16	0.25	0.05	0.27	**0.44**	0.30	0.39	**0.39**
King's College	**0.64**	**0.76**	0.12	0.41	0.12	0.09	0.06	−0.17
Old Hospital	0.27	**0.15**	0.16	0.00	0.35	**0.15**	**0.52**	0.12
Shop Façade	−0.28	−0.30	0.55	0.10	**0.58**	0.08	0.37	**0.12**
St Mary's Church	0.15	0.15	0.21	0.24	**0.42**	**0.32**	0.39	**0.32**
KTH Day 6 → 9	0.10	0.03	0.45	0.41	0.51	0.43	**0.62**	**0.52**
Average	0.12	0.12	0.31	**0.37**	0.39	0.32	**0.45**	0.35

6 Results and Discussion

We start our discussion by examining the behavior of an absolute pose regression network when its test distribution only partially overlaps with the training distribution. This is illustrated for our method in Fig. 1 and alongside other baselines in Fig. 3, where the network trained on sequence `kth_day_06` is tested on `kth_day_09` from Multi-Campus Dataset [28]. The test camera trajectory only partially overlaps with the training trajectory considering both position and orientation of the cameras. Additionally, as shown in Fig. 1, test images are captured under significantly different illumination conditions. We can see that this distribution shift results in generally large prediction errors across all methods, with median translation errors exceeding $20m$, as reported in Table 1. However, Fig. 1 and 3 also highlight that in overlapping stretches of the two trajectories the predictions are better aligned with the ground truth. It is critical for an uncertainty quantification method to distinguish the in- from out-of-distribution regions. Qualitatively, our method shows the highest degree of success at this, with its estimated uncertainty showing the strongest correlation with the predic-

tion error. Next, we quantitatively compare the correlation scores across different methods.

Figure 4 presents the scatter plot of quantified uncertainty against translation and rotation errors for all methods on `kth_day_09` test samples. The scattered points should ideally form a monotonically increasing function, corresponding to a Spearman rank correlation of +1. Starting from our method, we observe a generally increasing trend in uncertainty for small prediction errors. However, this pattern breaks for larger errors—uncertainty becomes *numb* to increases in error. We hypothesize that this occurs because, once critically far from the training distribution, the VAE model cannot differentiate between different test distributions. In the third and fourth columns of the same figure we filter out data points with translation errors larger than an empirical value of $50m$, which increases correlation coefficients from 0.62 and 0.52 to 0.75 and 0.67, showing the good performance of our method for test samples are not too far from the training distribution. Turning our attention to the results of other methods in Fig. 4, we see that other methods show generally lower correlations compared to ours. We also observe the same *numbing* phenomenon in their epistemic uncertainties, though even more severe, as it is harder to empirically filter out their insensitive data points. In the case of HR-APR we can see uncertainty values that saturate at a maximum dissimilarity for predicted poses that deviate significantly from the training samples. We note that while this strategy penalizes the monotonic relationship we evaluate, it can be a practical and cost-effective technique for outlier rejection.

Table 2 summarizes the correlation coefficients across different methods and environment. We see that the best performing method can differ between different scenes; BPN performs its best in outdoor scenes, while RVL and HR-APR generally perform better indoors. Our method, however, shows the most consistent performance across all datasets and has the highest average correlation between the quantified epistemic uncertainty and the prediction error. Importantly, our method achieves this while significantly outperforming the baselines methods in ambiguous scenarios with repetitive structures in terms of prediction accuracy, as seen in Table 1. In computing the median prediction errors, except for HR-APR that is deterministic, we draw 100 pose prediction samples per image observation and compute the error for the closest 10% of predicted samples to the ground truth. This is to handle ambiguous scenarios that result in multi-modal posterior distribution of camera poses given an image [43]. As evidenced by Table 1, on unambiguous datasets our method's accuracy is on par with the baselines.

As an outlook on future work, we reflect on the large variance of correlations across different scenes. Table 2 shows that, despite the improvements of our method, some scenes still exhibit correlations close to zero. This highlights the need to enhance the robustness of quantified epistemic uncertainty, potentially by incorporating temporal constraints. Another avenue for future research is addressing the *numbing* effect observed in Fig. 4. One possible approach is to

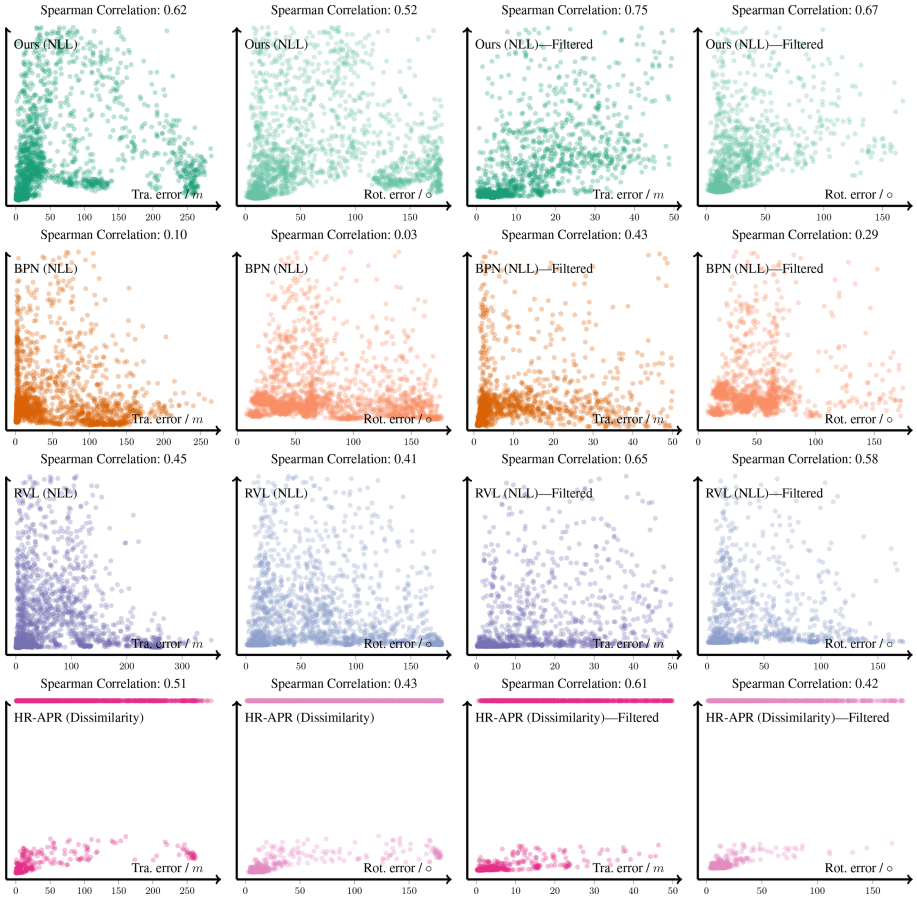

Fig. 4. The relationship between quantified epistemic uncertainty and prediction error across methods for our test sequence from Multi-Campus Dataset. The first two columns present scatter plots of all test samples, while the last two columns focus on samples with a translation error of less than $50m$. Since Spearman's rank correlation is invariant to affine changes, we scale and shift the uncertainty distribution in each plot so that their 10th and 90th percentiles align with those of the error distributions for better visualization, hence removing y-axis ticks. Our proposed approach demonstrates higher correlations between uncertainty and error, both with and without sample filtering.

frame this as an out-of-distribution binary classification problem, leveraging a method conceptually similar to HR-APR [21].

7 Conclusion

In this work, we address a key challenge in using absolute pose regression for visual relocalization: estimating the reliability of predictions for queries outside

the training distribution. We propose a method for quantifying epistemic uncertainty by estimating the likelihood of observations belonging to the training distribution. We leverage variational autoencoders for modeling this distribution and performing likelihood estimation, which acts as a unified framework for also modeling aleatoric uncertainty to handle observation ambiguities. Through extensive evaluation, we show the effectiveness of our approach in comparison to existing methods, and point out directions for future work.

Acknowledgements. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

1. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: CVPR, pp. 5297–5307 (2016)
2. Brachmann, E., Rother, C.: Learning less is more-6d camera localization via 3d surface regression. In: CVPR, pp. 4654–4662 (2018)
3. Brachmann, E., Rother, C.: Visual camera re-localization from RGB and RGB-D images using DSAC. TPAMI **44**(9), 5847–5865 (2021)
4. Brahmbhatt, S., Gu, J., Kim, K., Hays, J., Kautz, J.: Geometry-aware learning of maps for camera localization. In: CVPR, pp. 2616–2625 (2018)
5. Bürki, M., et al.: Vizard: reliable visual localization for autonomous vehicles in urban outdoor environments. In: 2019 IEEE Intelligent Vehicles Symposium (IV), pp. 1124–1130. IEEE (2019)
6. Chen, S., Li, X., Wang, Z., Prisacariu, V.A.: DFNet: enhance absolute pose regression with direct feature matching. In: ECCV, pp. 1–17. Springer (2022)
7. Chen, S., Wang, Z., Prisacariu, V.: Direct-PoseNet: absolute pose regression with photometric consistency. In: 3DV, pp. 1175–1185. IEEE (2021)
8. Cummins, M., Newman, P.: FAB-MAP: probabilistic localization and mapping in the space of appearance. IJRR **27**(6), 647–665 (2008)
9. Deng, H., Bui, M., Navab, N., Guibas, L., Ilic, S., Birdal, T.: Deep bingham networks: dealing with uncertainty and ambiguity in pose estimation. IJCV, 1–28 (2022)
10. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: ICML, pp. 1050–1059. PMLR (2016)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
12. Huang, Z., Xu, Y., Shi, J., Zhou, X., Bao, H., Zhang, G.: Prior guided dropout for robust visual localization in dynamic environments. In: ICCV, pp. 2791–2800 (2019)
13. Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K.M., Trulls, E.: Image matching across wide baselines: from paper to practice. IJCV **129**(2), 517–547 (2021)
14. Kendall, A., Cipolla, R.: Modelling uncertainty in deep learning for camera relocalization. In: ICRA, pp. 4762–4769 (2016)
15. Kendall, A., Cipolla, R.: Geometric loss functions for camera pose regression with deep learning. In: CVPR, pp. 5974–5983 (2017)

16. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? NeurIPS **30** (2017)
17. Kendall, A., Grimes, M., Cipolla, R.: PoseNet: a convolutional network for real-time 6-DOF camera relocalization. In: CVPR, pp. 2938–2946 (2015)
18. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: ICLR (2014)
19. Kingma, D.P., Welling, M., et al.: An introduction to variational autoencoders. Found. Trends® Mach. Learn. **12**(4), 307–392 (2019)
20. Liu, C., Chen, S., Zhao, Y., Huang, H., Prisacariu, V., Braud, T.: HR-APR: apr-agnostic framework with uncertainty estimation and hierarchical refinement for camera relocalisation. In: ICRA, pp. 8544–8550 (2024)
21. Liu, L., Li, H., Dai, Y.: Efficient global 2D-3D matching for camera localization in a large-scale 3D map. In: ICCV, pp. 2372–2381 (2017)
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
23. Melekhov, I., Ylioinas, J., Kannala, J., Rahtu, E.: Image-based localization using hourglass networks. In: ICCV Workshops, pp. 879–886 (2017)
24. Moreau, A., Piasco, N., Tsishkou, D., Stanciulescu, B., de La Fortelle, A.: CoordiNet: uncertainty-aware pose regressor for reliable vehicle localization. In: CVPR, pp. 2229–2238 (2022)
25. Moreau, A., Piasco, N., Tsishkou, D., Stanciulescu, B., de La Fortelle, A.: LENS: localization enhanced by NeRF synthesis. In: CoRL, pp. 1347–1356. PMLR (2022)
26. Naseer, T., Burgard, W.: Deep regression for monocular camera-based 6-DoF global localization in outdoor environments. In: IROS, pp. 1525–1530 (2017)
27. Ng, T., Lopez-Rodriguez, A., Balntas, V., Mikolajczyk, K.: Reassessing the limitations of CNN methods for camera pose regression. arXiv preprint arXiv:2108.07260 (2021)
28. Nguyen, T.M., et al.: MCD: diverse large-scale multi-campus dataset for robot perception. In: CVPR, pp. 22304–22313 (2024)
29. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) NeurIPS, pp. 8024–8035. Curran Associates, Inc. (2019)
30. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: robust hierarchical localization at large scale. In: CVPR, pp. 12716–12725 (2019)
31. Sattler, T., Leibe, B., Kobbelt, L.: Efficient & effective prioritized matching for large-scale image-based localization. TPAMI **39**(9), 1744–1756 (2016)
32. Sattler, T., Zhou, Q., Pollefeys, M., Leal-Taixe, L.: Understanding the limitations of CNN-based absolute camera pose regression. In: CVPR, pp. 3302–3312 (2019)
33. Schindler, G., Brown, M., Szeliski, R.: City-scale location recognition. In: CVPR, pp. 1–7. IEEE (2007)
34. Se, S., Lowe, D.G., Little, J.J.: Vision-based global localization and mapping for mobile robots. IEEE Trans. Rob. **21**(3), 364–375 (2005)
35. Shavit, Y., Ferens, R., Keller, Y.: Learning multi-scene absolute pose regression with transformers. In: ICCV, pp. 2733–2742 (2021)
36. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene coordinate regression forests for camera relocalization in RGB-D images. In: CVPR, pp. 2930–2937 (2013)
37. Toft, C., Maddern, W., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Pajdla, T., et al.: Long-term visual localization revisited. TPAMI **44**(4), 2074–2088 (2020)
38. Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. In: CVPR, pp. 1808–1817 (2015)

39. Walch, F., Hazirbas, C., Leal-Taixe, L., Sattler, T., Hilsenbeck, S., Cremers, D.: Image-based localization using LSTMs for structured feature correlation. In: ICCV, pp. 627–637 (2017)
40. Walker, J., Doersch, C., Gupta, A., Hebert, M.: An uncertain future: forecasting from static images using variational autoencoders. In: ECCV, pp. 835–851. Springer (2016)
41. Wang, B., Chen, C., Lu, C.X., Zhao, P., Trigoni, N., Markham, A.: AtLoc: attention guided camera localization. In: AAAI, vol. 34, pp. 10393–10401 (2020)
42. Zangeneh, F., Bruns, L., Dekel, A., Pieropan, A., Jensfelt, P.: A probabilistic framework for visual localization in ambiguous scenes. In: ICRA, pp. 3969–3975 (2023)
43. Zangeneh, F., Bruns, L., Dekel, A., Pieropan, A., Jensfelt, P.: Conditional variational autoencoders for probabilistic pose regression. In: IROS (2024)
44. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR, pp. 5745–5753 (2019)