2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Conditional Variational Autoencoders for Probabilistic Pose Regression

Fereidoon Zangeneh[1,2], Leonard Bruns[1], Amit Dekel[2], Alessandro Pieropan[2] and Patric Jensfelt[1]

***Abstract*— Robots rely on visual relocalization to estimate their pose from camera images when they lose track. One of the challenges in visual relocalization is repetitive structures in the operation environment of the robot. This calls for probabilistic methods that support multiple hypotheses for robot's pose. We propose such a probabilistic method to predict the posterior distribution of camera poses given an observed image. Our proposed training strategy results in a generative model of camera poses given an image, which can be used to draw samples from the pose posterior distribution. Our method is streamlined and well-founded in theory and outperforms existing methods on localization in presence of ambiguities.**

I. Introduction

Localization is one of the key capabilities that robots rely on for navigation. It is the task of estimating a robot's position and orientation in its operation environment from its sensor readings. Visual localization refers to the family of methods performing this task using image observations seen by an onboard camera. It comprises techniques for both frame-to-frame visual localization, where the relative pose between two camera views is estimated, as well as global visual relocalization, where the pose for a camera view is estimated with respect to the map. An effective global relocalization technique is essential for the operation of a robot when it does not have a prior estimation of its pose, such as at start-up, when it loses track, or in the case of a kidnapped robot.

The essence of global visual relocalization, hereafter referred to as visual relocalization, is maintaining a map representation that is most apt for retrieving the pose of a novel image observation. A variety of approaches have been explored to address this task, traditionally ranging from creating a database of past image observations and comparing new observations with them [1]–[3], to building a map of salient point features in the scene and performing structure-based point matching to estimate the camera pose [4]–[6]. More recently a trend of works explored storing representations of the scene in weights of a neural network that directly predicts the pose of a query image [7]–[9].

The majority of these developed methods focus on finding the best match for a query image, and there exist several works that achieve high localization accuracy when such a single best match exists [4, 10, 11]. However, a scenario that raises a challenge for general visual relocalization methods is when the operation environment is inherently ambiguous, that is due to its repetitive structures, distinct camera poses record the same observation. Examples of such ambiguities include stairs in a staircase, similar chairs in a meeting room as shown in Fig. 1, office doors in a corridor, or machine-fabricated panels of a ceiling. The existing probabilistic approaches that can accommodate multiple hypotheses for the camera pose either rely on predicting a mixture model [12] or learn a sampling-based mapping from the space of observed visual features to camera poses within the scene [13]. However, they require some level of prior knowledge about the number of modes present in the target probability distribution, rendering them more of ad hoc solutions rather than rigorous and general approaches.

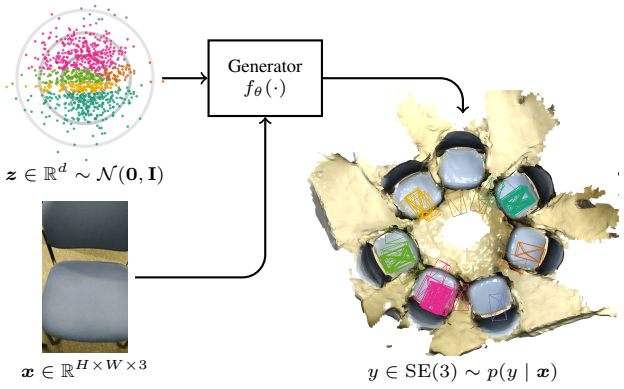


Fig. 1. Our proposed generative model takes in as input a query image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}$ as well as a random sample $\boldsymbol{z} \in \mathbb{R}^d$ drawn from the multivariate standard Gaussian distribution, and generates a sample $y \in \mathrm{SE}(3)$ from the posterior distribution of camera poses $p(y \mid \boldsymbol{x})$. Different regions of $\mathbb{R}^d$ are mapped to distinct modes in $\mathrm{SE}(3)$, color-coded for visualization.

In this work we propose a novel approach with theoretical grounding for learning the conditional distribution of camera poses in a scene given a query image. Our method learns the space of camera pose ambiguities based on the visual appearances viewed in the scene, solely from a set of image and camera pose label pairs. We make a conscious choice of following the learning-based pose regression paradigm, bearing in mind its limitations in accuracy and generalization [14] that call for targeted remedies [11]. We show that absolute pose regression provides a solid basis for our principled framework for probabilistic visual relocalization using conditional variational autoencoders.

In summary: (1) We propose a principled approach to probabilistic visual relocalization with a generative model that predicts samples from the posterior distribution of cam-

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

[1] Authors are with the division for Robotics, Perception and Learning, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden. {fzk,leonardb,patric}@kth.se

[2] Authors are with Univrses AB, SE-120 32 Stockholm, Sweden. {firstname.lastname}@univrses.com

era poses given an image. (2) We propose a training strategy for this generative model to learn the space of camera pose ambiguities in a scene using conditional variational autoencoders. (3) We perform a thorough evaluation to show that our method performs better than existing solutions for handling visual ambiguities.

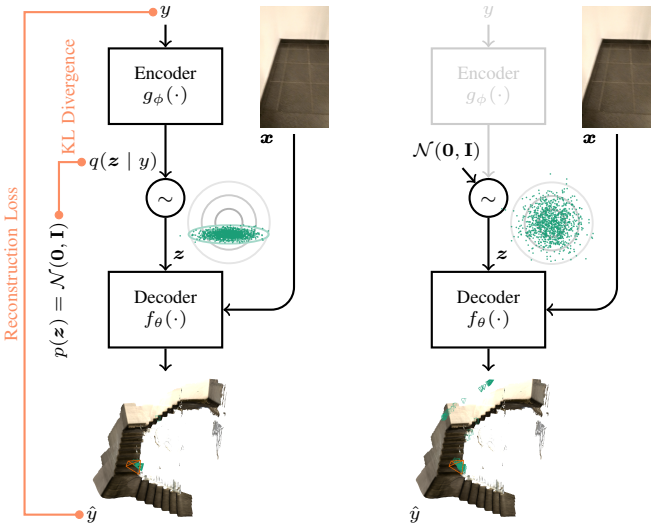
II. Related Work

A. Visual relocalization

Estimation of camera pose given an image is a long-standing problem in computer vision [15]–[17]. It is traditionally tackled by recognition-based [1, 2] and structure-based [4, 18] matching solutions, with the former group known for their efficiency and scalability, and the latter for higher accuracy of their solutions. Recently, learning-based regression methods have gained traction as a promising approach for visual relocalization. These methods train neural networks that directly solve for the pose of a camera in the scene given an image. They demand smaller memory and compute resources than the traditional methods and promise higher robustness in the face of illumination changes, motion blur and texture-less surfaces [8].

One variant of regression methods is scene coordinate regression, originally proposed for RGB-D camera pose estimation [19] and later extended to RGB data [20, 21]. These methods predict the scene's 3D point coordinates for image patches, from which the camera pose can be retrieved using a robust estimator. Although scene coordinate regression achieves similar accuracy levels as structure-based methods, end-to-end training of such pipelines for RGB data demands careful handling. A simpler approach of similar spirit is absolute pose regression, first proposed by Kendall et al. [7], where the camera pose is directly regressed from an image. This is achieved in a pipeline with a pretrained feature extractor, whose features are mapped to the absolute camera pose in the scene by a small multi-layer perceptron. Multiple improvements to the basic idea have been subsequently investigated, for example, the use of more carefully-crafted loss functions [22, 23] and alternative architectural choices [8, 24]–[26].

As discussed by Sattler et al. [14], absolute pose regression in its basic form is mainly comparable to pose approximation through image retrieval. Its generalization performance to novel views falls short in comparison to image retrieval and subsequent relative geometric pose estimation via feature-matching, and consistently underperforms compared to scene coordinate regression [21]. Various approaches have been proposed to alleviate this downside. One promising direction is to improve the generalization by synthesizing additional, novel, views from the training data. This idea has been employed for RGB-D data [27], and for RGB data via estimated depth maps [28] and NeRFs [11]. Other directions to improve the generalization include, for example, the use of equivariant features [29] and added photometric consistency constraints on the synthesized views of predictions [9, 30]. In this work we base our proposed method on the absolute pose regression paradigm for its easier end-to-end training



(a) Training (b) Inference

Fig. 2. (a) Our pose generative model is trained as the decoder in a conditional variational autoencoder pipeline reconstructing the ground-truth pose $y \in \mathrm{SE}(3)$ for an image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}$. The loss terms used in the learning objective are shown in orange. During training the latent posterior only partly overlaps with the latent prior, resulting in generated pose samples concentrated at the ground-truth pose. (b) At inference time latent samples are drawn from the prior distribution and mapped to distinct modes in $\mathrm{SE}(3)$. In the 3D rendering of the scene we can see that for the query image viewing an ambiguous landing at the staircase, output pose samples are concentrated at three modes looking at different, but visually similar landings, including the ground truth. Pose samples are shown by teal and the ground truth by orange camera frusta.

compared to scene coordinate regression. Additionally we incorporate data augmentation using NeRFs similar to [11] in one of our experiments. However, our proposed method specifically addresses the challenge of ambiguous scenes, making it orthogonal to the aforementioned advances.

B. Uncertainty estimation

A number of works propose estimating measures of uncertainty for the predictions. These works range from estimating epistemic uncertainty by training an ensemble of networks [31], to estimating homoscedastic [22] and heteroscedastic [32] aleatoric uncertainty by predicting a parametric distribution instead of a point prediction. Specifically focusing on multimodal distributions occurring for ambiguous images, Deng et al. [12] propose to parameterize the pose distribution as a mixture of joint Gaussian-Bingham distributions. Here, the Gaussian part describes the distribution of the position and the Bingham part describes the orientation distribution. To encourage multimodal distributions a winner-takes-all training scheme is used, in which only the mixture component closest to the target is supervised. This scheme has also been previously employed for object pose estimation [33]. With the goal of removing the explicit pose distribution parametrization, Zangeneh et al. [13] propose a variational inference framework in which an encoder predicts a latent distribution, which is subsequently decoded into a camera pose. Inspired by [12], a winners-take-all scheme is em-

ployed. These approaches require to provide the number of expected modes in advance; in [12] the number of mixture components, and in [13] the percentage of "winners" has to be specified. In this work, we propose a novel conditional variational autoencoder formulation, that does not include a comparable maximum or prior on the number of modes.

Conditional variational autoencoders have been used for other vision tasks to handle multimodal target distributions. In one of the first works that promoted the use of a conditional variational autoencoder framework, Sohn et al. [34] showcased generation of handwritten digits given a partial observation. Other works include, for example, forecasting dense pixel trajectories for static images [35] and image-to-image translation for domain transfer [36]. We take inspiration from these works to formulate a principled solution to handle ambiguous scenes in visual relocalization.

III. Method

For visual relocalization in presence of visual ambiguities, given an image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}$, we are interested in estimating its posterior distribution over all camera poses $p(y \mid \boldsymbol{x})$, where $y \in \mathrm{SE}(3)$. We outline our representation and approach for estimating this distribution in Section III-A. We then lay down the training scheme for our learned approach in Section III-B. We finally include a short discussion on the intuition behind our method in Section III-C.

A. Pose generative model

We propose to train a neural network $f_\theta(\cdot)$ that given an input image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}$, generates a sample $y \in \mathrm{SE}(3)$ from its pose posterior distribution. The source of randomness for this pose generator is the multivariate standard Gaussian distribution, from which a set of random samples $\mathcal{Z} = \{\boldsymbol{z}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \mid j = 1, \ldots, M\}$ can be drawn and fed as input to generate an arbitrarily-large set of pose samples $\mathcal{Y} = \{y_j = f_\theta(\boldsymbol{z}_j, \boldsymbol{x}) \mid \boldsymbol{z}_j \in \mathcal{Z}\}$ that represents $p(y \mid \boldsymbol{x})$. The generator network, depicted in Fig. 1, can be considered a learned random variable transformation of $\boldsymbol{z} \in \mathbb{R}^d$ to camera pose y, conditioned on the query image $\boldsymbol{x}$.

In order for the generative model to produce meaningful samples that represent $p(y \mid \boldsymbol{x})$, it must learn an appropriate mapping of $\mathbb{R}^d$ to $\mathrm{SE}(3)$, such that it transforms densities of $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to $p(y \mid \boldsymbol{x})$. In other words, $\mathbb{R}^d$ is a latent space, where different regions map to the various modes in $\mathrm{SE}(3)$ in the case of a multimodal $p(y \mid \boldsymbol{x})$. This means the generator relies on an organization of the latent space, such that, given an image, latent samples that generate pose samples around the same mode in $\mathrm{SE}(3)$ are clustered together, and all samples collectively are distributed according to $\mathcal{N}(\mathbf{0}, \mathbf{I})$. We propose to learn such an organization of the latent space for all observations that the generator can be conditioned on through a conditional variational autoencoder pipeline.

B. Training as a conditional variational autoencoder

We train the pose generative model as the conditional decoder of a variational autoencoder reconstructing pose samples, shown in Fig. 2(a). We refer the reader to [37, 38] for a thorough introduction to variational autoencoders, and include a short summary here for completeness.

1) Setting: The training process of the generative model assumes a dataset of training images with known camera poses taken within the scene $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \mid i = 1, \ldots, N\}$.

2) Learning-by-reconstruction: We frame the learning of the relation between camera poses and observations within the scene as a reconstruction task through the generative model. This reconstruction pipeline, shown in Fig. 2(a), consists of an inference network (encoder) $g_\phi(\cdot)$ with a Gaussian inference model, and the generative model (decoder) $f_\theta(\cdot)$. For an input pose y_i the encoder predicts the mean and covariance of the Gaussian posterior distribution in the latent space $q(\boldsymbol{z} \mid y_i)$, which we can easily draw samples from. The decoder, conditioned on the observed image $\boldsymbol{x}_i$, attempts to reconstruct the original pose y_i from a latent sample $\boldsymbol{z}_j \sim q(\boldsymbol{z} \mid y_i)$. This pipeline describes a variational autoencoder of camera poses conditioned on images. Following the general training scheme of variational autoencoders, the intractable true posterior distribution in the latent space $p(\boldsymbol{z} \mid y_i)$ is approximated by a class of known distributions (in this case Gaussian), and the per-pose inference of $q(\boldsymbol{z} \mid y_i)$ is amortized by training the encoder to directly predict the distribution parameters.

3) Optimization objective terms: The variational autoencoder pipeline is trained by maximizing the evidence lower bound (ELBO) [38] that consists of two terms: the Kullback-Leibler divergence $D_{\mathrm{KL}}(q(\boldsymbol{z} \mid y) \parallel p(\boldsymbol{z}))$ between the posterior $q(\boldsymbol{z} \mid y_i)$ and the prior distribution $p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ in the latent space, and the expected log-likelihood of reconstructed samples $\mathbb{E}_{q(\boldsymbol{z}|y)}[\log p(y \mid \boldsymbol{z}, \boldsymbol{x})]$. The latter is estimated by Monte Carlo simulation of the posterior $\boldsymbol{z}_j \sim q(\boldsymbol{z} \mid y_i)$ to predict $\hat{y}_{i,j} = f_\theta(\boldsymbol{z}_j, \boldsymbol{x}_i)$, and compute a reconstruction loss $d_{\text{pose}}(\hat{y}_{i,j}, y_i)$. The optimization objective then becomes

$$\min_{\theta,\phi} \sum_{\boldsymbol{x}_i, y_i \in \mathcal{D}} \Bigg[\beta \, D_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid y_i) \parallel p(\boldsymbol{z})) + \frac{1}{|\mathcal{Z}_i|} \sum_{\boldsymbol{z}_j \in \mathcal{Z}_i} d_{\text{pose}}(f_\theta(\boldsymbol{z}_j, \boldsymbol{x}_i), y_i) \Bigg], \tag{1}$$

where $\mathcal{Z}_i = \{\boldsymbol{z}_j \sim q_\phi(\boldsymbol{z} \mid y_i) \mid j = 1, \ldots, M\}$ is the set of latent posterior Monte Carlo samples, $|\mathcal{Z}_i|$ its cardinality, and β is the weight to balance the two terms.

C. Intuition

During the optimization laid out in (1) the encoder learns to organize the camera poses in the latent space, such that the predicted latent posteriors $q(\boldsymbol{z} \mid y)$ of camera poses that view similar visual appearances are sufficiently different from each other, while collectively conforming to the standard Gaussian prior. This means that at inference time, samples from the standard Gaussian prior overlap with the latent posteriors of all camera poses that viewed a similar visual appearance during training, and as a result are mapped to various modes in $\mathrm{SE}(3)$. The mapping of different regions in the latent space to various modes in $\mathrm{SE}(3)$ is shown in Fig. 1, and

the difference between training and inference time is further illustrated in Fig. 2. This clustering of camera poses given an image is learned without any prior on the number of modes or shape of the target distribution. This resolves the limitations of the existing works closest to our method, where [12] predicts a mixture model requiring the maximum number of modes to be explicitly set in advance, and [13] relies on a winners-take-all scheme to learn multimodal distributions, governed by a hyperparameter α that implicitly affects the number of modes that can be modeled.

IV. Implementation Details

A. Network architecture

The generative model (decoder) consists of a ResNet-18 backbone (with last layer set to identity) to extract 512-dimensional image features, as well as a multilayer perceptron. Image features and the d-dimensional latent features are each mapped by a linear layer (+ReLU) to the common dimensionality of 64, added, and fed as input to a fully connected network of five linear layers (+ReLU) of 128 features. This network predicts a 3-vector for the translation component together with a 6-vector for the rotation component of $\mathrm{SE}(3)$. The choice of rotation representation follows the work of Zhou et al [39] proposing a continuous representation for regression by neural networks.

The inference network (encoder), only used at training time, is a fully connected network with five linear layers (+ReLU) of 128 feature, which takes in a 9-dimensional pose sample (with the same representation as the generative model's output) and predicts the d-dimensional mean $\boldsymbol{\mu}$ and the log-variance $\log \boldsymbol{\sigma}^2$ that define the posterior distribution in the latent space $q(\boldsymbol{z} \mid y) = \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))$.

B. Training setup

We train the variational autoencoder pipeline using the Adam optimizer with a learning rate of 1×10^{-4} and a decoupled weight decay of 1×10^{-3} [40]. Following prior works [7, 12, 13] and for improved generalization with respect to lighting changes and motion blur we apply color and brightness jittering as well as Gaussian blur on the training images, followed by a random crop of 224×224 (after they are resized such that the smallest edge is 256 pixels). We define the reconstruction loss as

$$d_{\text{pose}}([\hat{\mathbf{R}} \mid \hat{\boldsymbol{t}}], [\mathbf{R} \mid \boldsymbol{t}]) = \lambda_r \|\hat{\mathbf{R}} - \mathbf{R}\|_{\mathrm{F}} + \lambda_t \|\hat{\boldsymbol{t}} - \boldsymbol{t}\|_2, \quad (2)$$

with λ_r and λ_t as balancing weights tuned to reflect the metric size of the scenes. We weigh the KL divergence term by $\beta = 0.3$ with a warm-up period of 4000 starting after 1000 iterations. We define the latent space to be 4-dimensional and train with a batch size of 16 in all experiments. At both training and test time we draw 1000 latent samples to represent the camera pose posterior for each image.

V. Experiments

A. Datasets and scenarios

We evaluate our method in three settings, two from state-of-the-art works for benchmarking [12, 13], extended by one synthetic scenario for a controlled analysis of our method.

1) Real-life ambiguous scenes: We evaluate our method on the Ambiguous Relocalization sequences [12] as well as a sequence of a moving camera facing a ceiling from [13], which captures a scenario of more severe ambiguity. We use these sequences to benchmark our method against the prior works. As the sequences in Ambiguous Relocalization dataset are short, for better generalization and following [11] we augment each training set by training a NeRF on the training samples [41] and synthesizing novel views uniformly-spaced along the camera trajectory, perturbed by Gaussian noise with $\sigma = 10\mathrm{cm}$ and 10°. We perform this for our method as well as the baseline [13].

2) Real-life unambiguous scenes: We also evaluate our method on the visual relocalization datasets 7-Scenes [19] and Cambridge Landmarks [7] to show its performance on general visual relocalization task in natural scenes.

3) Synthetic ambiguities: In order to examine the central thesis of our proposed method and decouple it from disturbances to the learning task—lighting changes, motion blur and dataset imbalance, we create a synthetic scenario, where a camera observes a scene made up of three distinct colors. The images are solid colors and are viewed in succession at a steady rate of camera movement. This creates a case of complete ambiguity to a learned network, without any seemingly benign but ultimately unambiguous speckles or details that tend to exist in real-life data. We use this synthetic sequence to showcase the functionality of our proposed method in a completely controlled and simplified environment.

B. Evaluation protocol

The true pose posterior for an ambiguous query image is by definition non-unimodal and has probability density also at regions distinct from the ground-truth pose label. For this reason, following [13] we measure the quality of a predicted pose posterior distribution by its recall. For a query image, we count its predicted pose posterior distribution as true positive if it contains sufficient density in the vicinity of the ground-truth pose label, and a false negative otherwise. In our sample-based representation of the posterior we assess the sufficiency of density by checking if at least a ratio γ of all samples are within a threshold of distance to the ground-truth pose. The appropriate value of γ depends on the degree of inherent ambiguity in the scene. Following [13] we use $\gamma = 0.1$ for Ambiguous Relocalization sequences and $\gamma = 0.05$ for the more ambiguous ceiling sequence.

Existing works on the task of visual relocalization commonly assess accuracy through median translation and orientation errors [7, 23, 31]. We adopt this metric to evaluate our method on unambiguous scenes. To do so we obtain point estimates for the predicted distributions by computing arithmetic and chordal L_2 [42] means of the translation and rotation components of their samples, respectively, followed by median error calculation across the test set.

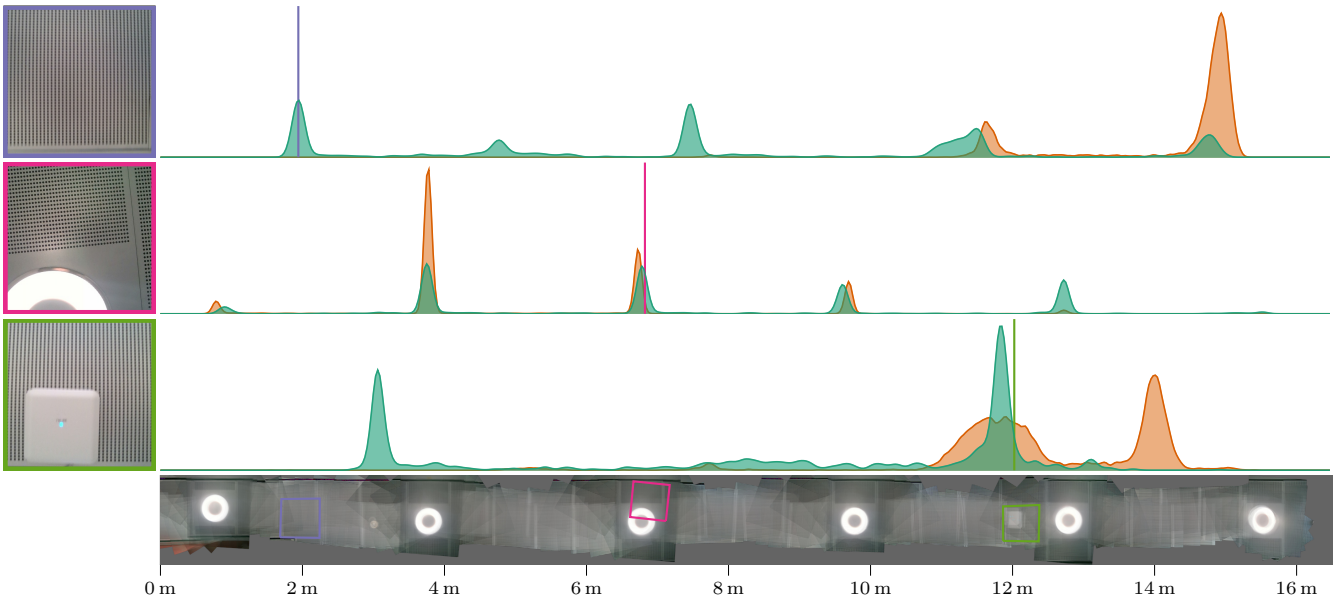


Fig. 3. Predicted marginal posterior distributions for the ceiling scene. Each row depicts an example, with the query image on the left, and predictions of our method in teal and VaPoR [13] in orange on the right. We represent each distribution with 1000 samples and show the marginalized distribution along the longitudinal translation axis. For visualization we show the results of kernel density estimation on the predicted samples. The ground-truth position in each example is shown by a vertical solid line, color-coded with the border of the query image and its projection on the ceiling. We can see that our method produces better predictions; in the first example VaPoR fails to assign density at the ground-truth position, and in the third example we deem the prediction of our method more coherent, predicting density at the position of the viewed WiFi router, as well as the smoke detector on the ceiling.

TABLE I

MEASURED RECALL IN AMBIGUOUS SCENES FOR THRESHOLDS $0.1\text{m}/10°$, $0.2\text{m}/15°$, $0.3\text{m}/20°$ (HIGHER IS BETTER)

#	MN [23]	BMDN [12]	VaPoR [13]	Ours
1	0.05/0.33/0.56	0.41/0.83/0.89	0.72/**1.00**/**1.00**	**0.79**/0.96/0.96
2	0.00/0.03/0.07	0.09/0.27/0.33	0.05/0.43/**0.61**	**0.15**/**0.45**/0.53
3	0.07/0.17/0.29	0.24/0.48/0.69	0.19/**0.61**/**0.78**	**0.29**/0.58/0.75
4	0.01/0.03/0.07	0.11/0.43/0.60	0.01/0.19/0.42	**0.15**/**0.48**/**0.62**
5	0.09/0.37/0.53	0.38/0.79/0.91	0.45/0.86/0.93	**0.52**/**0.92**/**0.98**
6	0.02/0.05/0.09	0.08/0.19/0.30	0.33/0.60/0.68	**0.53**/**0.73**/**0.81**

Scenes: (1) Blue Chairs (Fig. 1), (2) Meeting Table, (3) Staircase (Fig. 2), (4) Staircase Extended, (5) Seminar, (6) Ceiling (Fig. 3).

TABLE II

MEDIAN ERROR ($\text{m}/°$) IN UNAMBIGUOUS SCENES (LOWER IS BETTER)

Scene	MN* [23]	BPN* [31]	BMDN* [12]	VaPoR [13]	Ours
Chess	**0.08**/**3.25**	0.37/7.24	0.10/6.47	0.15/11.8	0.15/7.04
Fire	0.27/**11.7**	0.43/13.7	**0.26**/14.8	0.32/15.7	0.30/13.6
Heads	0.18/13.3	0.31/**12.0**	**0.13**/13.4	0.21/16.1	0.15/14.3
Office	**0.17**/**5.15**	0.48/8.04	0.19/9.73	0.24/12.1	0.24/8.99
Pumpkin	0.22/**4.02**	0.61/7.08	**0.20**/9.40	0.27/13.2	0.30/8.54
Kitchen	0.23/**4.93**	0.58/7.54	**0.19**/10.9	0.25/14.2	0.28/9.48
Stairs	**0.30**/**12.1**	0.48/13.1	0.34/14.1	0.30/16.2	0.36/14.5
College	1.07/**1.89**	1.74/4.06	1.51/2.14	1.07/4.27	**1.01**/3.74
Hospital	**1.94**/**3.91**	2.57/5.14	2.25/3.93	2.74/5.18	2.59/4.35
Façade	1.49/**4.22**	1.25/7.54	3.52/5.41	1.00/4.84	**0.98**/5.22
Church	2.00/**4.53**	2.11/8.38	2.16/5.99	1.86/7.13	**1.53**/7.94

* Results of MapNet, Bayesian PoseNet and BMDN are taken from [13].

VI. RESULTS AND DISCUSSION

A. Real-life ambiguous scenes

We compare our method to Deep Bingham Networks [12] (marked BMDN) and VaPoR [13], looking at the recall values of each across the ambiguous scenes. While capable of handling ambiguities in the scene, these two reference probabilistic pose regression methods, at least implicitly, rely on prior knowledge regarding the maximum number of modes in the target distribution. We evaluate BMDN with 10 mixture components accommodating a maximum of 10 modes, and VaPoR with the implicit parameter $\alpha = 0.05$ and a decoder depth of 5, giving both methods sufficient capacity to handle the ambiguities of the test scenes. We identified and rectified an implementation error in VaPoR's codebase, resulting in improved performance compared to their original recall values reported in [13]. In our evaluation we also include MapNet [23] (marked MN) as an effective end-to-end absolute pose regression method to highlight the need for probabilistic approaches for handling ambiguities.

We can see in Table I that our proposed method has an edge on other methods, while not assuming any prior knowledge about the ambiguities in the scene. Fig. 3 depicts three examples from the ceiling sequence, showcasing our method's more coherent predictions in comparison with VaPoR. Across different scenes we observed that our method often predicts distributions with sharper peaks compared to VaPoR, which tends to predict modes with wider spreads that are possibly favored by the adopted recall-based evaluation protocol. As also reported in [13], we observed that although BMDN's winner-takes-all strategy effectively positions mixture components, it struggles to predict the correct mixture weights required for generating coherent distributions.

B. Real-life unambiguous scenes

For completeness and to ensure the applicability of our method in a general visual relocalization setting, we re-

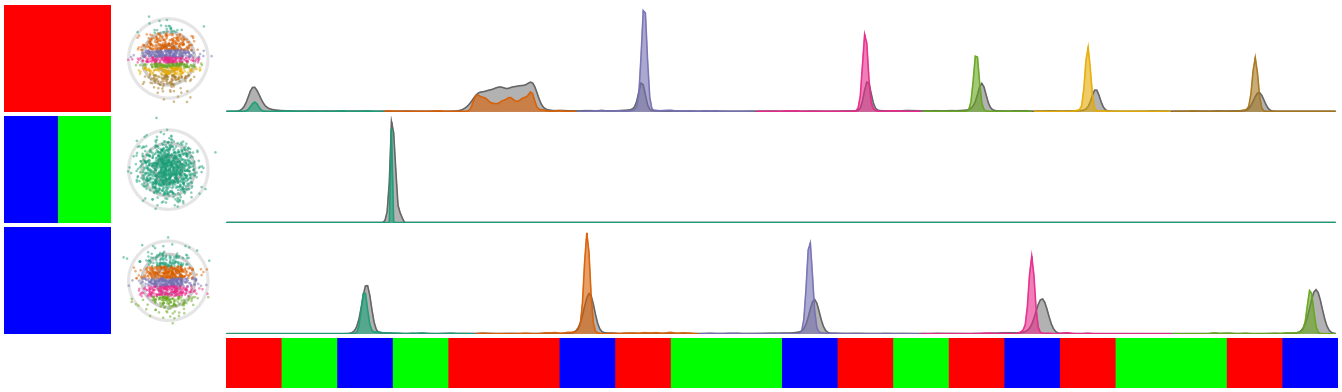


Fig. 4. Predicted marginal posterior distributions for our synthetic scene. Each row depicts an example, with the query image on the far left, followed by samples from the latent space, and predicted posterior distributions on the right. We obtain each distribution through a similar procedure as in Fig. 3, and color-code predictions of our method together with the latent samples according to the ground-truth modes of the posterior. Predictions of VaPoR [13] are shown in gray. We can see that our method can effectively learn the camera pose ambiguities for each query image, splitting the latent space into different regions mapping to distinct modes when the image is ambigous (first and third examples), while mapping the whole latent space to a single mode when the image is unambiguous (second example).

port the accuracy of our method in common unambiguous benchmark datasets in Table II. The table also includes the Bayesian PoseNet [31], another well-known absolute pose regression method. We see that our approach, even with the additional machinery of a sampling-based probabilistic method, achieves comparable performance to methods streamlined to predict a single accurate solution. This shows that our proposed framework does not impose an additional limitation on the capabilities of absolute pose regression.

C. Synthetic ambiguities

Fig. 4 illustrates our synthetic ambiguous sequence and highlights the relation between the latent space and the predicted posterior distribution for three example cases. In this controlled experiment, we know the true locations of the modes in the posterior distribution. So, we employ color-coding to associate the predicted posterior with the true modes, together with the samples drawn from the latent prior. We can see that the decoder can effectively learn the latent space of pose ambiguities given an image. That is, it splits the latent space into different regions depending on the ambiguity of the query image, and maps each to a distinct mode in $\mathrm{SE}(3)$.

D. Run-time analysis

We measure the time it takes to generate 1000 samples from the pose posterior distribution for a query image, corresponding to a forward pass of the generative model. With 100 repetitions, this on average takes $8.50 \pm 0.65\mathrm{ms}$ on CPU (Intel Core i9-13900KF), which means our method can run in real time.

VII. Future Work

We experimented with normalizing flows [43] for more flexible latent posterior shapes than Gaussian. However, we decided to exclude it as we did not observe consistent performance improvement across different scenes as a result of it. We hypothesize that in our setting the space of camera pose ambiguities could be modeled sufficiently well by low-dimensional Gaussian posteriors. However, future work could revisit the use of normalizing flows for likelihood estimation of pose samples [44]. Another promising direction is the use of diffusion models, which have shown remarkable generative capabilities in other tasks [45]–[47]. They can supplant the variational autoencoder setup for handling multimodal target distributions in visual relocalization.

VIII. Conclusion

In this work we revisit the problem of visual relocalization in the face of ambiguities and propose a novel approach with theoretical grounding and derived from first principles. At the core of our method is the learning of the space of camera pose ambiguities for image observations in a scene. We show that this can be materialized in a conditional variational autoencoder pipeline, yielding a generative model that given an image produces samples from its camera pose posterior distribution. Unlike the existing works, our proposed approach does not assume any prior knowledge about the shape of the target distribution. We perform an extensive evaluation to examine the working of our method and compare it to existing methods, showing that our approach has a performance edge on other works in ambiguous scenes.

References

[1] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1808–1817.

[2] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.

[3] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 141–14 152.

[4] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1744–1756, 2016.

[5] L. Liu, H. Li, and Y. Dai, "Efficient global 2D-3D matching for camera localization in a large-scale 3D map," in *Proceedings of the International Conference on Computer Vision*, 2017, pp. 2372–2381.

[6] T. Sattler, B. Leibe, and L. Kobbelt, "Improving image-based localization by active correspondence search," in *Proceedings of the European Conference on Computer Vision*. Springer, 2012, pp. 752–765.

[7] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2938–2946.

[8] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using LSTMs for structured feature correlation," in *Proceedings of the International Conference on Computer Vision*, 2017, pp. 627–637.

[9] S. Chen, Z. Wang, and V. Prisacariu, "Direct-PoseNet: Absolute pose regression with photometric consistency," in *Proceedings of the International Conference on 3D Vision*. IEEE, 2021, pp. 1175–1185.

[10] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 716–12 725.

[11] A. Moreau, N. Piasco, D. Tsishkou, B. Stanciulescu, and A. de La Fortelle, "LENS: Localization enhanced by NeRF synthesis," in *Conference on Robot Learning*. PMLR, 2022, pp. 1347–1356.

[12] H. Deng, M. Bui, N. Navab, L. Guibas, S. Ilic, and T. Birdal, "Deep bingham networks: Dealing with uncertainty and ambiguity in pose estimation," *International Journal of Computer Vision*, pp. 1–28, 2022.

[13] F. Zangeneh, L. Bruns, A. Dekel, A. Pieropan, and P. Jensfelt, "A probabilistic framework for visual localization in ambiguous scenes," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2023, pp. 3969–3975.

[14] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixe, "Understanding the limitations of CNN-based absolute camera pose regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3302–3312.

[15] S. Se, D. G. Lowe, and J. J. Little, "Vision-based global localization and mapping for mobile robots," *IEEE Transactions on Robotics*, vol. 21, no. 3, pp. 364–375, 2005.

[16] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–7.

[17] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[18] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "InLoc: Indoor visual localization with dense matching and view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7199–7209.

[19] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in rgb-d images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2930–2937.

[20] E. Brachmann and C. Rother, "Learning less is more-6d camera localization via 3d surface regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4654–4662.

[21] ——, "Visual camera re-localization from RGB and RGB-D images using DSAC," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5847–5865, 2021.

[22] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5974–5983.

[23] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz, "Geometry-aware learning of maps for camera localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2616–2625.

[24] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu, "Image-based localization using hourglass networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 879–886.

[25] B. Wang, C. Chen, C. X. Lu, P. Zhao, N. Trigoni, and A. Markham, "AtLoc: Attention guided camera localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 06, 2020, pp. 10 393–10 401.

[26] Y. Shavit, R. Ferens, and Y. Keller, "Learning multi-scene absolute pose regression with transformers," in *Proceedings of the International Conference on Computer Vision*, 2021, pp. 2733–2742.

[27] T. Naseer and W. Burgard, "Deep regression for monocular camera-based 6-DoF global localization in outdoor environments," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017, pp. 1525–1530.

[28] T. Ng, A. Lopez-Rodriguez, V. Balntas, and K. Mikolajczyk, "Reassessing the limitations of cnn methods for camera pose regression," *arXiv preprint arXiv:2108.07260*, 2021.

[29] M. A. Musallam, V. Gaudilliere, M. O. Del Castillo, K. Al Ismaeil, and D. Aouada, "Leveraging equivariant features for absolute pose regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6876–6886.

[30] S. Chen, X. Li, Z. Wang, and V. A. Prisacariu, "DFNet: Enhance absolute pose regression with direct feature matching," in *Proceedings of the European Conference on Computer Vision*. Springer, 2022, pp. 1–17.

[31] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2016, pp. 4762–4769.

[32] A. Moreau, N. Piasco, D. Tsishkou, B. Stanciulescu, and A. de La Fortelle, "CoordiNet: uncertainty-aware pose regressor for reliable vehicle localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2229–2238.

[33] J. Fu, Q. Huang, K. Doherty, Y. Wang, and J. J. Leonard, "A multi-hypothesis approach to pose ambiguity in object-based slam," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2021, pp. 7639–7646.

[34] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[35] J. Walker, C. Doersch, A. Gupta, and M. Hebert, "An uncertain future: Forecasting from static images using variational autoencoders," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 835–851.

[36] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[37] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[38] D. P. Kingma, M. Welling *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[39] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5745–5753.

[40] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[41] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 102:1–102:15, Jul. 2022. [Online]. Available: https://doi.org/10.1145/3528223.3530127

[42] R. Hartley, J. Trumpf, Y. Dai, and H. Li, "Rotation averaging," *International Journal of Computer Vision*, vol. 103, no. 3, pp. 267–305, 2013.

[43] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2015, pp. 1530–1538.

[44] N. Kolotouros, G. Pavlakos, D. Jayaraman, and K. Daniilidis, "Probabilistic modeling for human mesh recovery," in *Proceedings of the International Conference on Computer Vision*, 2021, pp. 11 605–11 614.

[45] J. Gong, L. G. Foo, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, "DiffPose: Toward more reliable 3d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 041–13 051.

[46] J. Zhang, M. Wu, and H. Dong, "Generative category-level object pose estimation via diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[47] R. Shrestha, B. Koju, A. Bhusal, D. P. Paudel, and F. Rameau, "CaLDiff: Camera localization in NeRF via pose diffusion," *arXiv preprint arXiv:2312.15242*, 2023.