2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

A Probabilistic Framework for Visual Localization in Ambiguous Scenes

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

***Abstract*— Visual localization allows autonomous robots to relocalize when losing track of their pose by matching their current observation with past ones. However, ambiguous scenes pose a challenge for such systems, as repetitive structures can be viewed from many distinct, equally likely camera poses, which means it is not sufficient to produce a single best pose hypothesis. In this work, we propose a probabilistic framework that for a given image predicts the arbitrarily shaped posterior distribution of its camera pose. We do this via a novel formulation of camera pose regression using variational inference, which allows sampling from the predicted distribution. Our method outperforms existing methods on localization in ambiguous scenes. We open-source our approach and share our recorded data sequence at github.com/efreidun/vapor.**

I. Introduction

Visual localization is the task of inferring the ego pose of a camera from its image. It enables mobile robots to localize themselves in an environment, which is crucial for their navigation. Regardless of the paradigm that is followed to solve this task, the proposed methods revolve around detection of visual features that are unique to different regions of the environment and the camera poses that view them. Some methods do this by retrieving the most similar image to a query image from a database of images previously collected in the scene [1], [2], [3], [4]; some establish point correspondences between the salient features of the query image and a pre-built 3D feature map, and use projective geometry relations to estimate the camera pose [5], [6], [7], [8], [9]; and some delegate this estimation problem to end-to-end learning-based solutions that regress the camera pose from what it views [10], [11], [12], [13], [14].

As long as there are unique identifying features in the images, there exist numerous solutions that can accurately estimate the camera pose [6], [15]. However, the same cannot be said when the scene is ambiguous [16], that is, when it contains distinct regions that are visually indistinguishable. Examples of this include identical doors, identical chairs arranged around a table, or the flights of stairs in a staircase, as illustrated in Fig. 1. A desired solution in these cases is one that produces multiple pose hypotheses, capturing the repetitive patterns of the scene, rather than attempting to produce a single best hypothesis. This calls for a multi-hypothesis localization framework, which we address in this work. We focus on inference of the camera pose distribution from a single image, and refer to the rich literature on robot localization for how to accumulate evidence and maintain such a distribution over time [17], [18], [19].

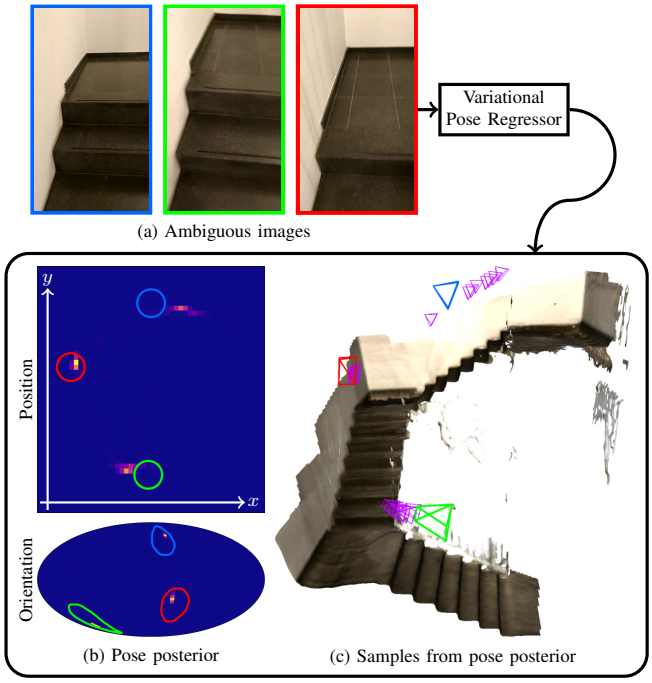


Fig. 1. (a) Visually similar images taken from three different flights of stairs, (b) camera pose distribution predicted for the right image, and (c) samples drawn from this posterior. The distribution is visualized by a position heatmap on the xy-plane (marginalizing height) and an elliptic orientation heatmap[2]. We show the drawn samples in a 3D reconstruction of the scene by small camera frusta in purple. The ground-truth camera poses are shown by color-coded circles in (b) and camera frusta in (c).

We propose a probabilistic framework that allows inferring the posterior distribution over camera poses for a given image. We represent this distribution by an arbitrary number of samples drawn from it, which in theory can model distributions with any number of modes and of any shape. Samples from this distribution can be used in downstream tasks, such as motion planning or active localization. We formulate our solution following the paradigm of end-to-end camera pose regression, and employ variational inference [21], [22] to model the visual features of images used for localization. We show that camera pose regression, despite its limitations in generalization and accuracy compared to structure-based methods [23], when combined with variational inference

* This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Authors thank Thien-Minh Nguyen for his help in recording and obtaining ground-truth poses for the new image sequence.

[1] Authors are with the division of 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-118 26 Stockholm, Sweden. {firstname.lastname}@univrses.com

[2]We use the Mollweide projection for the surface of the 2-sphere component of $SO(3)$ obtained through Hopf-fibration (marginalizing the fibers), inspired by Murphy et al. [20].

gives rise to a simple, yet powerful solution for pose posterior prediction from an observed image.

We summarize our contributions as the following: (1) We lay out a novel formulation of camera pose regression using variational inference, which allows sampling from an arbitrarily shaped pose distribution for a given image. (2) We propose a novel sampling-based Winners-Take-All optimization scheme, which allows learning multimodal distributions. (3) We record a sequence of real-world camera images capturing a case of severe visual ambiguity for evaluation of localization solutions. (4) We show that our formulation outperforms existing methods on ambiguous scenes.

II. Related work

Regression-based approaches aim to solve the pose estimation problem in a single step by finding a function that directly maps an image to its pose, promising improved performance in feature-less environments or under motion blur [12]. In early work, Shotton et al. [24] proposed regressing 3D scene coordinates for each pixel in an image. In combination with depth data, this allows robust estimation of the 6D pose of the camera by employing RANSAC with Kabsch's algorithm [25]. The first end-to-end approach for image-based pose regression was PoseNet proposed by Kendall et al. [10]. Specifically, they proposed to train a deep neural network to directly regress the 6D camera pose from the image features extracted by a pre-trained backbone.

Following this early work, various improvements orthogonal to our work have subsequently been proposed. Naseer and Burgard [26] showed that RGB-D data can be exploited to generate additional views from the limited training images to improve performance. Recently, Ng et al. [27] and Moreau et al. [15] extended this idea to RGB data. Other works propose to use additional information often available in robotic applications [11], [13].

More closely related to our work, several works investigate how to model uncertainty for pose regressors. In [28], the authors apply Bayesian deep learning to PoseNet. This allows one to gauge the uncertainty in the prediction, although the ability to learn more complicated distributions remains limited, as noted by [16]. While [28] focused on epistemic uncertainty, Kendall and Cipolla [29] considered homoscedastic aleatoric uncertainty by modifying the loss, and Moreau et al. [30] modeled heteroscedastic aleatoric uncertainty instead by predicting an uncertainty measure.

Deng et al. [16] further extend the idea of representing uncertainty by predicting a mixture of multiple unimodal distributions. In principle, this allows the network to correctly predict multiple modes for ambiguous queries. A downside to this mixture-based approach is the difficulty of picking the correct number of modes, and training the network so that it actually predicts different modes. To handle the latter issue, the authors propose a Winner-Takes-All scheme that only gives supervision to the best predicted mode. Our work follows a similar idea, but instead of employing a mixture model with a fixed number of components, we follow a variational approach, which, in principle, can learn to produce arbitrarily shaped pose distributions.

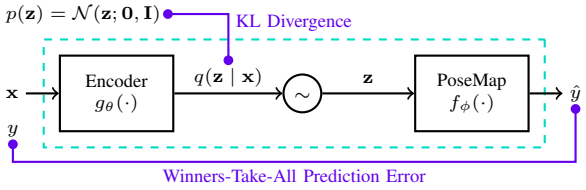


Fig. 2. Our pipeline for inference of the camera pose distribution for an image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ with ground-truth pose label $y \in \mathrm{SE}(3)$. We can simulate the posterior distribution $p(y \mid \mathbf{x})$ by drawing samples $\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x}), \mathbf{z} \in \mathbb{R}^d$ and applying the mapping $f_\phi(\mathbf{z})$ to get $\hat{y} \sim p(y \mid \mathbf{x}), \hat{y} \in \mathrm{SE}(3)$. The loss terms used in the learning objective are shown in purple, where the Winners-Take-All strategy confines the minimization of prediction error to the subset of samples $\hat{y}$ that are in the neighborhood of y.

In another related line of work, Murphy et al. [20] build on the recent success of neural fields [31] by employing an MLP that predicts the probability density for a given rotation, allowing representation of arbitrary distributions in $\mathrm{SO}(3)$. Our work also aims to learn arbitrary distributions, but we propose a sampling-based approach, in which a sample from a latent space is transformed to a pose in $\mathrm{SE}(3)$. This simplifies inference, as it does not require dense querying of the support to find the modes of the distribution; instead, our approach allows direct sampling from it.

III. Method

We propose to perform visual localization for an image in two steps: (1) infer a distribution in the latent space capturing the visual features that are useful for localization within the scene; (2) perform a random variable transformation to obtain a distribution of camera poses for the query image. Fig. 2 visualizes our proposed pipeline.

A. Formulation

Let $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ be a color image taken from camera pose $y \in \mathrm{SE}(3)$. In localization, where the scene is known beforehand, one can in theory infer the posterior distribution of visual features as seen in the observed image $p(\mathbf{z} \mid \mathbf{x})$. Here, $\mathbf{z} \in \mathbb{R}^d$ is the latent variable corresponding to the visual features that the scene comprises. With this definition of the latent variable, visually similar images result in similar posterior distributions in the latent space, even if the images are taken from distinct camera poses, as in ambiguous scenes.

Having full knowledge of the scene, the posterior distribution of visual features should contain the information needed to infer the posterior distribution of camera poses given the observed image $p(y \mid \mathbf{x})$. This can be formulated as a transformation of densities from visual features in $\mathbb{R}^d$ to camera pose in $\mathrm{SE}(3)$, which can be achieved by applying a deterministic mapping $f : \mathbb{R}^d \to \mathrm{SE}(3)$ to samples drawn from the posterior distribution in the latent space: $y = f(\mathbf{z}), \quad \mathbf{z} \sim p(\mathbf{z} \mid \mathbf{x})$.

B. Modeling via learning

In the proposed formulation, there are two scene-dependent operations that model the scene for the purpose

of visual localization, namely the inference of the posterior distribution in the latent visual features' space $p(\mathbf{z} \mid \cdot)$, and the mapping to camera pose $f(\cdot)$. We parameterize these in the weights of two deep neural networks and learn them from data samples collected from the scene. We refer to the two networks as *Encoder* $g_\theta(\cdot)$ and *PoseMap* $f_\phi(\cdot)$, parameterized by θ and ϕ, respectively.

Encoder $g_\theta(\cdot)$ is an inference network with a Gaussian inference model that for an input image $\mathbf{x}$ outputs $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\sigma} \in \mathbb{R}^d$ defining the posterior distribution $q(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))$ in the latent space. This follows the variational principle, where an unknown posterior distribution is modeled by optimizing the parameters of a convenient family of distributions such as Gaussians to best resemble the true posterior. Akin to Variational Auto-Encoders (VAEs) [21], [22], we amortize this per-image optimization at inference time by optimizing the *Encoder* weights at training time to directly predict the distribution parameters.

PoseMap $f_\phi(\cdot)$ is a fully connected network that, for an input sample from the latent space $\mathbf{z}$, outputs a camera pose y. This means that the posterior distribution of the camera pose $p(y \mid \mathbf{x})$ can be approximated by simulating the inferred posterior distribution in the latent space $\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})$ via reparameterization trick and passing the drawn samples through the mapping $y = f_\phi(\mathbf{z})$ to obtain samples $y \sim p(y \mid \mathbf{x})$. The output of the network y comprises a translation vector $\mathbf{t} \in \mathbb{R}^3$ and a 6D representation for rotation $\mathbf{r} \in \mathbb{R}^6$. The rotation parameterization choice is the continuous representation for rotations in 3D introduced by Zhou et al. [32], where a rotation matrix is retrieved from the 6D representation following a Gram-Schmidt-like process.

C. Learning scheme

The network weights θ, ϕ that represent a scene are learned from a dataset of images and camera poses $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid i = 1, \ldots, N\}$ collected in that scene. For this, we lay out an optimization scheme that enables learning multimodal pose distributions as is desired in ambiguous scenes.

1) Objective terms: **Prediction error** measures the quality of a predicted pose $\hat{y} \in \mathrm{SE}(3)$ against its ground truth y. We define the prediction error as the weighted sum of a translation error term defined on $\mathbb{R}^3$ and a rotation error term defined on $\mathrm{SO}(3)$. The translation error is the Euclidean distance between the translation components $\hat{\mathbf{t}}, \mathbf{t} \in \mathbb{R}^3$ of predicted and ground-truth poses. For the rotation error, we opt for the chordal distance between the rotation components $\hat{\mathbf{R}}, \mathbf{R} \in \mathrm{SO}(3)$, as its gradient is well-defined everywhere over its domain, as opposed to, for example, the geodesic distance. The prediction error is thus defined as

$$d_{\text{pose}}(\hat{y}, y) = \lambda_t \|\hat{\mathbf{t}} - \mathbf{t}\|_2 + \lambda_r \|\hat{\mathbf{R}} - \mathbf{R}\|_{\mathrm{F}}, \tag{1}$$

where λ_t and λ_r are tunable constants, balancing the scales of the two terms.

Kullback–Leibler divergence $D_{\mathrm{KL}}(q(\mathbf{z} \mid \mathbf{x}) \parallel p(\mathbf{z}))$ measures how different an inferred latent posterior distribution $q(\mathbf{z} \mid \mathbf{x})$ is from a prior distribution defined on the latent variable $p(\mathbf{z})$. This is an integral part of the variational principle, which together with the prediction error forms the evidence lower bound (ELBO) optimized in variational approaches. As is common practice, we assume a standard Gaussian prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$ for its simplicity in computing the KL divergence.

2) Evidence lower bound (ELBO): In variational approaches, the ELBO objective that is typically maximized is a combination of negative KL divergence and expected log-likelihood of predictions $\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p_\phi(y \mid \mathbf{z})]$. The latter expectation is generally computed by Monte Carlo simulation of $q(\mathbf{z} \mid \mathbf{x})$. With our choice of pose prediction error, the variational optimization objective can be written as

$$\min_{\theta,\phi} \sum_{\mathbf{x}_i, y_i \in \mathcal{D}} \Bigg[D_{\mathrm{KL}}(q(\mathbf{z} \mid \mathbf{x}_i) \parallel p(\mathbf{z})) + \frac{1}{|\mathcal{Z}_i|} \sum_{\mathbf{z}_j \in \mathcal{Z}_i} d_{\text{pose}}(f_\phi(\mathbf{z}_j), y_i) \Bigg], \tag{2}$$

where $\mathcal{Z}_i = \{\mathbf{z}_j \sim q(\mathbf{z} \mid \mathbf{x}_i) \mid j = 1, \ldots, M\}$ is the Monte Carlo sample set and $|\mathcal{Z}_i|$ its cardinality.

We argue that minimizing this objective, and specifically the expected prediction error, is counterproductive in our setting, where the camera pose posterior $p(y \mid \mathbf{z})$ can be multimodal in ambiguous scenes. In such scenarios, two visually similar images $\mathbf{x}_i$ and $\mathbf{x}_j$ $(i \neq j)$ are encoded to similar latent posterior distributions $p(\mathbf{z} \mid \mathbf{x}_i)$ and $p(\mathbf{z} \mid \mathbf{x}_j)$. However, these images can be taken from two distinct poses y_i and y_j in the scene, in which case the true posterior distributions of the camera pose $p(y \mid \mathbf{x}_i)$ and $p(y \mid \mathbf{x}_j)$ are both bimodal. Minimizing the expected prediction error results in a compromised solution in the form of a unimodal inferred distribution between the two true modes. We propose a modification of the expected error term to address this.

3) Winners-Take-All optimization: We propose to confine the computed mean prediction error to a subset of Monte Carlo samples $\hat{\mathcal{Z}}_i \subseteq \mathcal{Z}_i$, whose image through the mapping $f_\phi(\cdot)$ is within a certain distance δ of the true mode y_i, that is, $\hat{\mathcal{Z}}_i = \{\mathbf{z}_j \in \mathcal{Z}_i \mid d_{\text{pose}}(f_\phi(\mathbf{z}_j), y_i) < \delta\}$. This ensures that pose samples can concentrate around individual modes during optimization without influence from other modes. However, the true posterior is unknown and different modes can have different shapes, rendering the choice of δ non-trivial. Moreover, random initialization of the parameters θ and ϕ does not guarantee that there will be pose samples within any δ distance of the modes at the start of the optimization. This calls for an adaptive selection of δ at every iteration and for every mode.

At every iteration and for a ground-truth pose y_i we pick $\delta_{i,\alpha}$ as the radius of the smallest ball centered at y_i containing a fraction α of samples in $\mathcal{Z}_i$. In other words, our adaptive $\delta_{i,\alpha}$, defined as

$$\delta_{i,\alpha} = \inf \left\{ \delta \in \mathbb{R}_+ \;\middle|\; \left|\hat{\mathcal{Z}}_i\right| = \lfloor \alpha \cdot |\mathcal{Z}_i| \rfloor \right\}, \tag{3}$$

results in minimizing the prediction error for only the closest fraction α of Monte Carlo samples per ground-truth pose y_i.

Our proposed optimization objective is

$$\min_{\theta,\phi} \sum_{\mathbf{x}_i, y_i \in \mathcal{D}} \Bigg[\beta \; D_{\mathrm{KL}}(q(\mathbf{z} \mid \mathbf{x}_i) \parallel p(\mathbf{z})) + \frac{1}{|\hat{\mathcal{Z}}_{i,\alpha}|} \sum_{\mathbf{z}_j \in \hat{\mathcal{Z}}_{i,\alpha}} d_{\mathrm{pose}}(f_\phi(\mathbf{z}_j), y_i) \Bigg], \tag{4}$$

where $\hat{\mathcal{Z}}_{i,\alpha} = \{\mathbf{z}_j \in \mathcal{Z}_i \mid d_{\mathrm{pose}}(f_\phi(\mathbf{z}_j), y_i) < \delta_{i,\alpha}\}$. α and β are tunable constants, the latter being the balancing weight for the KL divergence term.

This is in spirit similar to the Winner-Takes-All multi-hypothesis optimization scheme used for learning mixture models, where the closest mixture component is optimized per label [16], [33]. However, our proposed solution is in a different setting, as we represent posteriors by samples instead of mixture models. We therefore refer to our method as Winners-Take-All to acknowledge this similarity, while reflecting the fact that it is used for optimizing sample sets rather than individual mixture components.

IV. Experiments, Results and Discussion

A. Implementation details

We implement our method using the PyTorch library [34]. We use ResNet-18 [35] as the backbone of the *Encoder* to extract 2048-dimensional feature vectors, followed by a linear layer to predict d-dimensional $\boldsymbol{\mu}$ and $\log \boldsymbol{\sigma}^2$ vectors for the latent posterior. The *PoseMap* is implemented with a fully connected network taking the input vector through the dimensionality transformation $d \rightarrow 128(\rightarrow 128)_{\times n_{\text{layers}}} \rightarrow 3 + 6$ with ReLU activations in-between. The minimum number of hidden layers n_{layers} depends on the complexity of the target pose distributions in the scene. In nearly all tested scenes we achieved favorable performance with as few as $n_{\text{layers}} = 3$, which, unless otherwise stated, is used across all experiments. The final layer corresponds to the prediction of translation and rotation vectors, where the former goes through a sigmoid activation, followed by a fixed affine transformation that shifts and scales the predictions to the metric ranges of the scene.

We train our networks using Adam optimizer [36] with initial learning rate of 1×10^{-4} and an exponential learning rate decay of 0.8, applied every $n_{\text{lr-decay}}$ epochs for 10 occurences. Following the pose regression literature, we first resize each image such that its smallest edge is 256, then randomly crop 224×224 regions for input to the *Encoder*. We also augment the data with color/brightness jittering and Gaussian blur to account for lighting changes and motion blur between images. Unless otherwise stated, we let $\alpha = 0.20$, $\beta = 0.01$, use a $d = 16$-dimensional latent space, and represent distributions with 1000 Monte Carlo samples in all experiments, since we found this to produce good predictions in our setting. Other hyperparameters are reported in Table I, tuned to reflect the number of images and metric scales of different datasets, which range from small indoor to large outdoor scenes. Note that we found these settings without a major hyperparameter search, and one may improve the performance by a thorough search of the optimal hyperparameters.

TABLE I
Hyperparameters used in training

Dataset	λ_t	λ_r	Batch Size	# Epochs	$n_{\text{lr-decay}}$
7-Scenes [24]	5	10	64	100	10
Cambridge Land. [10]	5	100	64	500	50
Ambiguous Reloc. [16]	5	2	4	500	50
Ceiling	5	2	4	2000	50
Synthetic	5	2	4	500	50

B. Datasets and metrics

We evaluate our method on the Ambiguous Relocalization dataset [16] as an existing benchmark with real-world image sequences of ambiguous environments. For each scene in the dataset there are separate training and test image sequences recorded from their own unique camera trajectories, but with generally similar views. We found that despite the apparent ambiguity to the human eye, a large fraction of frames in this dataset contain unique identifying features, which an expressive feature detector can infer the pose from. This results in unimodal predicted posteriors for a large number of frames, which hinders the evaluation of a method's capability in forming multimodal distributions. To address this, we complement the dataset by recording a new real-world sequence of a ceiling with machine-fabricated panels, capturing a case of severe visual ambiguity. We record the training and test sequences with a calibrated LiDAR-IMU-camera rig, and obtain ground-truth camera poses using MILIOM [37]. We also render image sequences of two synthetic scenes from 3D Warehouse[3], which contain symmetries by design, and use them to investigate our method in a controlled setting.

We use recall as the metric to evaluate pose distributions in ambiguous scenes. For a query image, we draw samples from its posterior distribution, and consider it a true positive if at least a fraction γ of the samples are within a distance of the ground-truth pose (and a false negative otherwise). We argue that for a distribution with well-separated equally likely modes, setting γ inversely proportional to the number of modes gives an estimate of whether the distribution contains sufficient density around the ground-truth pose. For simplicity, we report recall with $\gamma = 0.1$ for all tested scenes except for the ceiling scene that has a richer set of ambiguities, where we use $\gamma = 0.05$.

To validate the performance of our method as a general pose regressor on unambiguous scenes, we evaluate it on the visual localization benchmarks 7-Scenes [24] and Cambridge Landmarks [10]. As is commonly reported by pose regression works, we use median error for evaluation on these datasets. We obtain a point prediction from the Monte Carlo samples of each predicted distribution using the arithmetic and chordal L_2 [38] means for translation and orientation, respectively. The median of this estimate's error compared to the ground-truth pose is reported across each scene.

[3]https://3dwarehouse.sketchup.com/

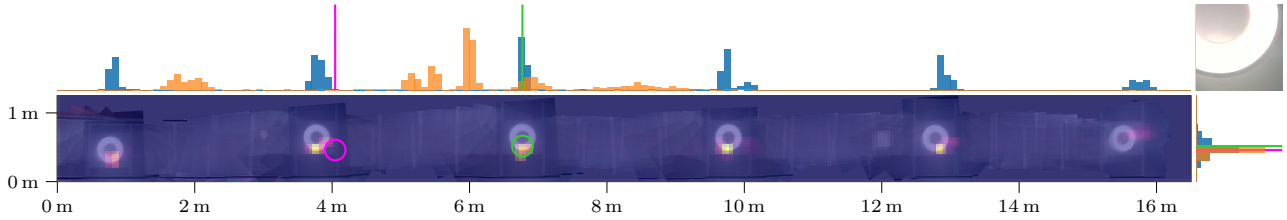


Fig. 3. Marginal posterior distributions along x-axis (top left) and y-axis (bottom right) predicted by our method (), by Bingham MDN [16] (), the prediction by MapMet [13] (), and the ground truth () for a query image (top right) from the ceiling scene. The heatmap shows the 2D histogram predicted by our method overlaid on top of stitched images of the scene. Note that our method successfully captures all six modes of the distribution while MapNet only predicts a single estimate at a wrong location, and Bingham MDN method assigns large probabilities in visually dissimilar locations.

TABLE II

MEASURED RECALL IN AMBIGUOUS SCENES (HIGHER IS BETTER)

Scene	Threshold	PN [10]	MN [13]	BMDN [16]	Abl.	Ours
Blue Chairs	0.1m/10°	0.08	0.05	0.41	0.32	**0.45**
	0.2m/15°	0.40	0.33	0.83	0.89	**0.99**
	0.3m/20°	0.56	0.46	0.89	0.97	**1.00**
Meeting Table	0.1m/10°	0.00	0.00	**0.09**	0.03	0.06
	0.2m/15°	0.02	0.03	0.27	0.24	**0.35**
	0.3m/20°	0.02	0.07	0.33	0.34	**0.43**
Staircase	0.1m/10°	0.00	0.07	0.24	0.12	**0.30**
	0.2m/15°	0.01	0.17	0.48	0.44	**0.62**
	0.3m/20°	0.01	0.29	0.69	0.63	**0.72**
Staircase Ext.	0.1m/10°	0.00	0.01	0.11	0.03	**0.12**
	0.2m/15°	0.00	0.03	0.43	0.24	**0.53**
	0.3m/20°	0.01	0.07	0.60	0.44	**0.71**
Seminar Room	0.1m/10°	0.00	0.09	0.38	0.17	**0.43**
	0.2m/15°	0.02	0.37	0.79	0.53	**0.90**
	0.3m/20°	0.10	0.53	0.91	0.80	**0.97**
Ceiling†	0.1m/10°	0.00	0.02	0.08	0.00	**0.09**
	0.2m/15°	0.03	0.05	0.19	0.02	**0.31**
	0.3m/20°	0.06	0.09	0.30	0.04	**0.44**

† We train the independently recorded ceiling scene with $\alpha = 0.05$ and $n_{\text{layers}} = 9$, reflecting the richer presence of ambiguities in the scene.

TABLE III

MEDIAN ERROR (M / °) IN UNAMBIGUOUS SCENES (LOWER IS BETTER)

Scene	PN [10]	MN† [13]	BPN [28]	BMDN [16]	Ours
Chess	0.32/8.12	**0.08/3.25**	0.37/7.24	0.10/6.47	0.17/6.90
Fire	0.47/14.4	0.27/**11.7**	0.43/13.7	**0.26**/14.8	0.30/14.1
Heads	0.29/**12.0**	0.18/13.3	0.31/**12.0**	**0.13**/13.4	0.17/14.5
Office	0.48/7.68	**0.17/5.15**	0.48/8.04	0.19/9.73	0.24/9.30
Pumpkin	0.47/8.42	0.22/**4.02**	0.61/7.08	**0.20**/9.40	0.30/8.33
Kitchen	0.59/8.64	0.23/**4.93**	0.58/7.54	**0.19**/10.9	0.26/10.2
Stairs	0.47/13.8	**0.30/12.1**	0.48/13.1	0.34/14.1	0.47/15.5
College	1.92/5.40	**1.07/1.89**	1.74/4.06	1.51/2.14	1.65/2.88
Street	3.67/6.50	–	**2.96/6.00**	16.3/25.2	17.2/23.8
Hospital	2.31/5.38	**1.94/3.91**	2.57/5.14	2.25/3.93	2.06/4.33
Façade	1.46/8.08	1.49/**4.22**	1.25/7.54	3.52/5.41	**1.02**/6.03
Church	2.65/8.48	2.00/**4.53**	2.11/8.38	2.16/5.99	**1.80**/5.90

† Results of MapNet on Cambridge Landmarks taken from Sattler et al. [23].

C. Evaluation on benchmark datasets

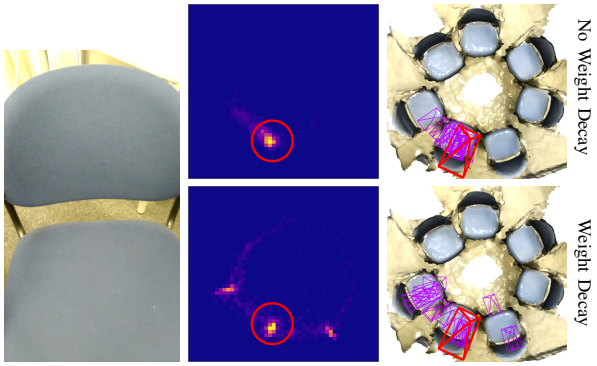
We report the results on the ambiguous scenes in Table II. We can see that our method, outperforms Bingham MDN [16] as the method closest to ours that predicts a distribution of poses aimed at localization in ambiguous scenes. Following their setting, we considered their approach with 10 and 50 components in their mixture model, and evaluated the metric based on samples drawn from them. As the 10-component setting consistently performed better, we report its results as a representative in the table (marked BMDN). Fig. 3 shows an example of the predicted posterior given a query image from the ceiling scene, where we can see posterior predicted by our method better captures the ambiguous structure of the scene. We also evaluate PoseNet [10] and its Bayesian variant [28], as well as MapNet [13]. However, we see that these single estimate methods fail to achieve comparable performance on the ambiguous scenes. To our surprise, vanilla PoseNet performed comparatively better than Bayesian PoseNet, so we include its results as representative (marked PN) alongside MapNet (marked MN).

In order to investigate whether our method's improved performance stems from our novel formulation with variational inference, we perform an ablation, in which we modify our pipeline to produce a single pose for an input image. We remove the KL divergence term from the objective, modify the *Encoder* to predict a single point, and obtain a single pose prediction by passing the encoder's prediction through *PoseMap*. All else equal, we evaluate this ablative variant of our method that is in principle very similar to PoseNet. We can see in Table II that this variant, marked *Abl.*, while performing better than PoseNet due to its more recent feature extractor network, falls short of the unablated variant, validating the merit of our proposed formulation.

For completeness, we report our results on the unambiguous 7-Scenes and Cambridge Landmarks datasets in Table III. We include results of PoseNet and MapNet as single pose regressor baselines, and Bayesian PoseNet and Bingham MDN as methods that, in principle, can predict multimodal distributions. While our method does not perform the best, it is not far from the top-performers. This experiment merely serves as a sanity check of our approach's performance in a minimal pipeline, without any particular mechanism aimed at improving accuracy in unambiguous scenes. As seen in Table II, the better performing methods on unambiguous scenes show poor performance on ambiguous scenes, which is the problem that our method targets to solve. An interesting direction for future work is to apply our proposed formulation, aimed at handling ambiguous scenes, in tandem with techniques for improved unambiguous pose regression.

D. A closer look

Fig. 4 (top) shows the predicted distribution by our method for an example query image from the Ambiguous Relocalization dataset. Although the scene, made up of identical chairs, is arguably ambiguous to the human eye, we can see that



(a) Query image (b) Position posterior (c) Samples

Fig. 4. Posteriors predicted by our method in full capacity (top) and in a constrained learning mode by L_2 weight decay of $\lambda = 0.1$ (bottom). The latter captures multiple modes whereas the former mode predicts a single mode at the correct pose.

the predicted posterior identifies and concentrates its density around the correct pose. We hypothesize that a sufficiently expressive *Encoder* can distinguish a seemingly ambiguous image taken in real life by its smallest of details, such as the chair's background in this example. However, a less expressive *Encoder* for the data is unable to learn every detail and can give in to the ambiguities. We test this hypothesis by adding a penalty term on the L_2 norm of the *Encoder* weights during training. We can see in Fig. 4 (bottom) that this setting results in the predicted posterior assigning probabilities to poses viewing two additional chairs. We argue that when there exists a domain gap between the training data and the operation conditions, it is desirable for the model to trade off confidence in predictions for better generalization, which can be achieved via deliberate learning constraints. We leave the study of such learning constraints to future work.

We study the effect of α in the Winners-Take-All optimization scheme in two synthetic scenes, where the camera circles around a round table with four legs, resulting in four modes in the pose distribution of an image, as well as a rectangular dinner table that results in bimodal distributions. We report the statistics over 10 training runs for the $0.1\text{m}/10^\circ$ recall evaluated at the end of training with different α values in Fig. 6. We can see that in these scenes the highest recall is achieved with α in a range of values greater than zero but less than $1/(\#\text{modes})$. Fig. 5 shows the predicted camera position posterior for three choices of α. We can see that a too large α, as discussed in Section III-C.2, results in a compromised posterior, and a too small α predicts close-to-uniform densities across the span of the training data. We hypothesize that α must be smaller than $1/(\#\text{modes})$ for the Winners-Take-All optimization to converge and capture all modes in the distribution, and must be sufficiently larger than zero to overcome the noise as a result of mini-batch optimization. There is a trade-off between training speed and the quality of the learned distribution within this range of α values, as a smaller α results in optimization of fewer samples at every iteration, hence a slower training, but is less

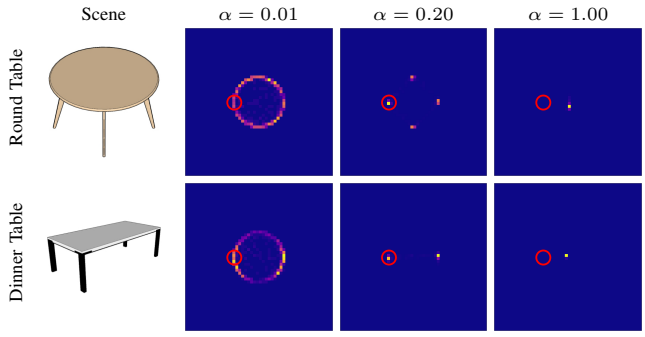
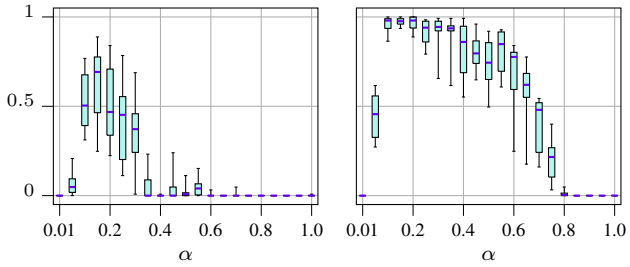


Fig. 5. Predicted position posterior on xy-plane for various choices of α in the synthetic scenes. The round table and dinner table have four and two modes in their true distributions, respectively. The optimization successfully converges to predict the correct modes when $0 \ll \alpha < 1/(\#\text{modes})$. For example $\alpha = 0.20$ produces good predictions for both scenes.



(a) Round Table (b) Dinner Table

Fig. 6. Measured recall with threshold $0.1\text{m}/10^\circ$ for various choices of α in Winners-Take-All optimization of the synthetic scenes. For each choice of α, the statistics over 10 training runs is shown by a box extending from the lower to upper quartile recall values, a purple line at the median recall, and whiskers that extend to the minimum and maximum values.

susceptible to the noise induced by Monte Carlo sampling. We leave the study of finding the optimal α to future work.

E. Run-time evaluation

We measure the time taken for a forward pass of one query image through our pipeline for 1000 Monte Carlo samples, on a desktop computer with an Intel Core i7-8700K CPU and an NVIDIA GeForce GTX 1080 Ti GPU. We repeat each measurement 100 times and we find that a forward pass on average takes 14.88 ± 0.75 ms on CPU and 2.26 ± 0.10 ms on GPU, that is, our pipeline can run in real time.

V. Conclusion

In this work, we propose a novel formulation of camera pose regression with variational inference to address the task of visual localization in ambiguous scenes. Our approach allows learning the distribution of latent visual features present in the scene that are useful for localization, and hence learning the distribution of camera poses given an image and sampling from it. Through a series of experiments, we show that this variational approach enables prediction of high-quality multimodal pose distributions that is required for localization in ambiguous scenes, and performs better than existing methods that attempt to directly learn and represent the camera pose distribution with mixture models. We also propose directions for further investigation of our method.

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] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2017, pp. 3223–3230.

[4] 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.

[5] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua, "Worldwide pose estimation using 3D point clouds," in *Proceedings of the European Conference on Computer Vision*. Springer, 2012, pp. 15–29.

[6] 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.

[7] 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.

[8] P.-E. Sarlin, A. Unagar, M. Larsson, H. Germain, C. Toft, V. Larsson, M. Pollefeys, V. Lepetit, L. Hammarstrand, F. Kahl *et al.*, "Back to the feature: Learning robust camera localization from pixels to pose," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3247–3257.

[9] 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.

[10] 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.

[11] R. Clark, S. Wang, A. Markham, N. Trigoni, and H. Wen, "VidLoc: A deep spatio-temporal model for 6-DoF video-clip relocalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6856–6864.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] D. Fox, S. Thrun, W. Burgard, and F. Dellaert, "Particle filters for mobile robot localization," in *Sequential Monte Carlo methods in practice*. Springer, 2001, pp. 401–428.

[18] P. Jensfelt and S. Kristensen, "Active global localization for a mobile robot using multiple hypothesis tracking," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 5, pp. 748–760, 2001.

[19] D. Fox, "KLD-Sampling: Adaptive particle filters," in *Advances in Neural Information Processing Systems*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds., vol. 14. MIT Press, 2001.

[20] K. A. Murphy, C. Esteves, V. Jampani, S. Ramalingam, and A. Makadia, "Implicit-PDF: Non-parametric representation of probability distributions on the rotation manifold," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 7882–7893.

[21] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proceedings of the International Conference on Learning Representations*, 2014.

[22] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2014, pp. 1278–1286.

[23] 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.

[24] 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.

[25] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 32, no. 5, pp. 922–923, 1976.

[26] 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.

[27] 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.

[28] 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.

[29] ——, "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.

[30] 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.

[31] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. Sitzmann, and S. Sridhar, "Neural fields in visual computing and beyond," *Computer Graphics Forum*, vol. 41, no. 2, pp. 641–676, 2022.

[32] 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.

[33] O. Makansi, E. Ilg, O. Cicek, and T. Brox, "Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7144–7153.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, 2015.

[37] T.-M. Nguyen, S. Yuan, M. Cao, L. Yang, T. H. Nguyen, and L. Xie, "Miliom: Tightly coupled multi-input lidar-inertia odometry and mapping," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5573–5580, 2021.

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