arXiv:2301.02086v1 [cs.CV] 5 Jan 2023

歧义场景中视觉定位的概率框架

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

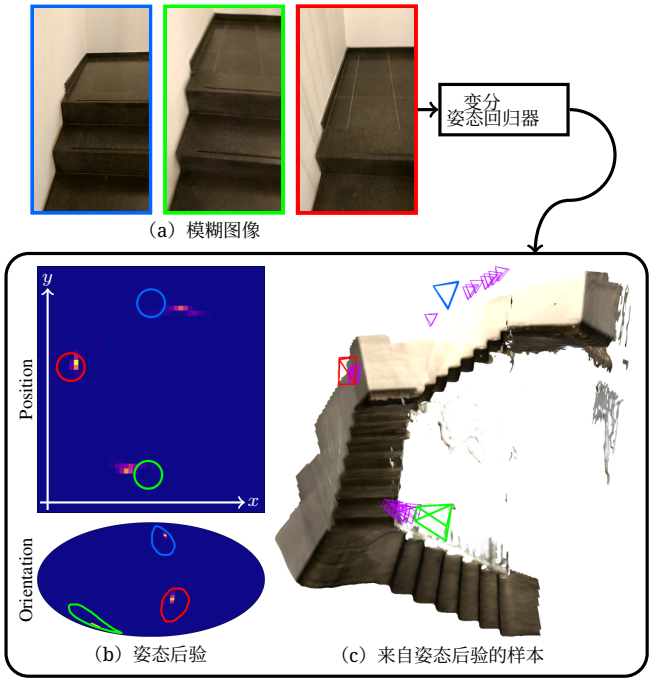
摘要——视觉定位使自主机器人在失去位姿追踪后，能够通过将当前观测与过去的观测进行匹配来重新定位。然而，模糊场景给这类系统带来了挑战，因为重复结构可能对应许多不同但同样可能的相机位姿，因此仅输出单个位姿假设并不足够。本文提出一种概率框架，可针对给定图像预测其相机位姿形状任意的后验分布。我们通过一种运用变分推断的相机位姿回归新形式化来实现这一点，从而能够从预测分布中采样。我们的方法在模糊场景定位方面优于现有方法。代码和数据将发布于 github.com/efreidun/vapor。

I. 引言

视觉定位是根据相机图像推断其自我位姿的任务。它使移动机器人能够在环境中定位自身，这对其导航至关重要。无论采用何种范式来解决这一任务，所提出的方法都围绕着检测环境不同区域中具有独特性的视觉特征，以及能够观测到这些特征的相机位姿展开。有些方法通过从先前在场景中采集的图像数据库中检索与查询图像最相似的图像来实现这一点 [1], [2], [3], [4]；有些方法则在查询图像的显著特征与预先构建的三维特征地图之间建立点对应关系，并利用射影几何关系估计相机位姿 [5], [6], [7],[8], [9]；还有一些方法将这一估计问题交由端到端的学习方法解决，根据相机所观测到的内容回归相机位姿 [10], [11], [12], [13], [14]。

只要图像中存在独特的识别特征，就有许多方法能够准确估计相机位姿 [6], [15]。然而，当场景存在歧义时，情况就不同了 [16], ，也就是说，场景中包含多个视觉上无法区分的不同区域。例如，完全相同的门、围绕桌子摆放的相同椅子，或楼梯间中的楼梯梯段，如图 1 所示。在这些情况下，理想的解决方案应生成多个位姿假设，以捕捉场景中的重复模式，而不是试图给出单一的最佳假设。为此，需要一个多假设定位框架，我们将在本文中探讨这一问题

* 本研究部分得到了由 Knut and Alice Wallenberg Foundation 资助的 Wallenberg AI, Autonomous Systems and Software Program (WASP) 的支持。作者感谢 Thien-Minh Nguyen 协助录制新图像序列并获取其真实位姿。[1] 作者隶属于瑞典皇家理工学院机器人、感知与学习系，地址：瑞典斯德哥尔摩 SE-10044。{fzk,leonardb,patric}@kth.se [2] 作者隶属于 Univrses AB，地址：瑞典斯德哥尔摩 SE-11826。{firstname.lastname}@univrses.com



（a）模糊图像

（b）姿态后验

（c）来自姿态后验的样本

图1。（a）从三个不同楼梯梯段拍摄的视觉上相似的图像，（b）为右侧图像预测的相机位姿分布，以及（c）从该后验分布中抽取的样本。该分布通过 xy 平面上的位置热图（边缘化高度）和椭圆方向热图[2]进行可视化。我们在场景的三维重建中用紫色的小型相机视锥展示所抽取的样本。（b）中的真实相机位姿以颜色编码的圆圈表示，（c）中的真实相机位姿以相机视锥表示。

相关工作。我们聚焦于从单张图像推断相机位姿分布，并参考机器人定位领域丰富的研究成果，了解如何随时间累积证据并维护此类分布 [17], [18], [19]。

我们提出了一种概率框架，用于推断给定图像对应的相机位姿后验分布。我们通过从该分布中采样得到任意数量的样本来表示这一分布；从理论上讲，这种表示可以对具有任意数量众数和任意形状的分布进行建模。该分布中的样本可用于下游任务，例如运动规划或主动定位。我们遵循端到端相机姿态回归的范式来构建解决方案，并采用变分推断 [21], [22] 对用于定位的图像视觉特征进行建模。我们证明，尽管相较于基于结构的方法 [23], ，相机位姿回归在泛化能力和准确性方面存在局限，但将其与变分推断相结合

[2]受 Murphy 等人 [20] 的启发，我们使用莫尔威德投影表示通过霍普夫纤维化从 SO(3) 中得到的二维球面分量的表面（对纤维进行边缘化）。

能够从观测图像中实现一种简单而强大的姿态后验预测方案。

我们将贡献总结如下：(1) 我们提出了一种使用变分推断进行相机位姿回归的新颖形式化方法，使其能够从给定图像对应的任意形状位姿分布中进行采样。(2) 我们提出了一种新颖的基于采样的赢家通吃优化方案，能够学习多模态分布。(3) 我们记录了一段真实世界的相机图像序列，其中捕捉了一个视觉模糊性严重的案例，用于评估定位方案。(4) 我们表明，在模糊场景中，我们的形式化方法优于现有方法。

II. 相关工作

基于回归的方法旨在通过寻找一个将图像直接映射到其位姿的函数，在单步内解决位姿估计问题，从而有望在无特征环境或存在运动模糊的情况下提升性能 [12]。在早期工作中，Shotton等人 [24]提出了回归图像中每个像素对应的三维场景坐标。结合深度数据，通过采用结合Kabsch算法的随机抽样一致性方法，可以稳健地估计相机的六自由度位姿 [25]。首个端到端的基于图像的位姿回归方法是Kendall等人提出的PoseNet [10]。具体而言，他们提出训练一个深度神经网络，直接根据预训练骨干网络提取的图像特征回归六自由度相机位姿。

在这项早期工作的基础上，随后又提出了多种与我们工作正交的改进方法。Naseer 和 Burgard [26] 展示了如何利用 RGB-D 数据，从有限的训练图像中生成额外视图，以提升性能。最近，Ng 等人 [27] 和 Moreau 等人 [15] 将这一思路扩展到了 RGB 数据。其他研究则提出利用机器人应用中通常可获得的额外信息 [11], [13]。

与我们工作关系更密切的是，若干研究探讨了如何为位姿回归器建模不确定性。在[28], 中，作者将贝叶斯深度学习应用于 PoseNet。这样便可以评估预测的不确定性，不过正如 [16]所指出的，其学习更复杂分布的能力仍然有限。 [28] 侧重于认知不确定性，而 Kendall 和 Cipolla [29]通过修改损失函数来考虑同方差偶然不确定性；Moreau 等人 [30] 则通过预测不确定性度量来建模异方差偶然不确定性。

Deng 等人 [16] 进一步拓展了通过预测多个单峰分布的混合来表示不确定性的思路。原则上，这使网络能够针对模糊查询正确预测多个众数。这种基于混合分布的方法的缺点在于，难以确定合适的众数数量，也难以训练网络使其真正预测出不同的众数。为解决后一问题，作者提出了一种赢家通吃方案，只对预测效果最佳的众数提供监督信号。我们的工作遵循类似思路，但我们不采用具有固定数量分量的混合模型，而是

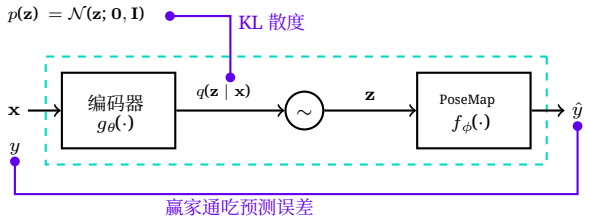


图 2. 我们用于推断图像 $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ 的相机位姿分布的流程，其真实位姿标签为 $y \in \mathrm{SE}(3)$。通过抽取样本 $\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x}), \mathbf{z} \in \mathbb{R}^d$，并应用映射 $f_\phi(\mathbf{z})$，即可模拟后验分布 $p(y \mid \mathbf{x})$，从而得到$\hat{y} \sim p(y \mid \mathbf{x}), \hat{y} \in \mathrm{SE}(3)$。学习目标中使用的损失项以紫色显示。

采用变分方法，原则上能够学习生成形状任意的位姿分布。

另一项相关研究中，Murphy 等人 [20] 基于神经场 [31] 近期取得的成功，采用多层感知机来预测给定旋转的概率密度，从而能够表示 SO(3) 上的任意分布。我们的工作同样旨在学习任意分布，但提出了一种基于采样的方法：将潜在空间中的样本变换为 SE(3) 中的位姿。这简化了推断过程，因为无需对支持集进行密集查询来寻找分布的众数；相反，我们的方法可以直接从分布中采样。

III. 方法

我们提出通过两个步骤对图像进行视觉定位：(1) 在潜在空间中推断一个分布，以捕获场景中有助于定位的视觉特征；(2) 执行随机变量变换，以获得查询图像的相机位姿分布。图 2 展示了我们提出的流程。

A. 形式化

设 $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ 为从相机位姿 $y \in$SE(3)处拍摄的彩色图像。在定位中，场景已知，因此理论上可以推断观测图像 $p(\mathbf{z} \mid \mathbf{x})$中视觉特征的后验分布。这里，$\mathbf{z} \in \mathbb{R}^d$ 是对应于场景所包含视觉特征的潜变量。按照此潜变量的定义，即使图像来自不同的相机位姿，视觉上相似的图像也会在潜在空间中产生相似的后验分布，例如在模糊场景中。

在完全了解场景的情况下，视觉特征的后验分布应包含根据观测图像 $p(y \mid \mathbf{x})$推断相机位姿后验分布所需的信息。这可以表述为将视觉特征 $\mathbb{R}^d$ 中的密度变换为 SE(3) 中相机位姿密度的问题；对此，可将确定性映射 $f : \mathbb{R}^d \to \mathrm{SE}(3)$ 应用于从潜在空间中的后验分布抽取的样本：$y = f(\mathbf{z}), \mathbf{z} \sim p(\mathbf{z} \mid \mathbf{x})$。

B. 通过学习进行建模

在所提出的形式化中，有两个依赖场景的操作用于对场景进行建模，以实现

视觉定位，即在潜在视觉特征空间 $p(\mathbf{z} \mid \cdot)$ 中推断后验分布，以及映射到相机位姿 $f(\cdot)$。我们将这两项操作参数化为两个深度神经网络的权重，并使用从场景中收集的数据样本进行训练。我们将这两个网络分别称为 编码器 $g_\theta(\cdot)$ 和 *PoseMap* $f_\phi(\cdot)$，其参数分别为 θ 和 ϕ。

编码器 $g_\theta(\cdot)$是一个推理网络，其高斯推理模型针对输入图像$\mathbf{x}$ 输出 $\boldsymbol{\mu} \in \mathbb{R}^d$和 $\boldsymbol{\sigma} \in \mathbb{R}^d$ ，用于定义潜在空间中的后验分布 $q(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\mathbf{z};\ \boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))$。这遵循变分原理：通过优化一类便于处理的分布（如高斯分布）的参数，使其尽可能逼近真实后验分布，从而对未知后验分布进行建模。与变分自编码器（VAEs）类似， [21][22], 我们通过在训练时优化编码器 的权重，使其能够直接预测分布参数，从而在推理时摊销每张图像对应的优化过程。

PoseMap $f_\phi(\cdot)$是一个全连接网络，它针对来自潜在空间$\mathbf{z}$的输入样本输出一个相机位姿 y。这意味着，可以通过重参数化技巧模拟潜在空间$\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})$中的推断后验分布，并将抽取的样本传入映射 $y = f_\phi(\mathbf{z})$，从而获得样本$y \sim p(y \mid \mathbf{x})$，以此近似相机位姿的后验分布 $p(y \mid \mathbf{x})$。网络 y 的输出由平移向量$\mathbf{t} \in \mathbb{R}^3$ 和旋转的六维表示$\mathbf{r} \in \mathbb{R}^6$组成。旋转参数化采用 Zhou 等人 [32], 提出的三维旋转连续表示，并通过类似格拉姆–施密特的过程，从六维表示中恢复旋转矩阵。

C. 学习方案

表示场景的网络权重 θ，ϕ 是通过该场景中收集的图像和相机位姿数据集 $\mathcal{D} = \{(\mathbf{x}_i,\ y_i) \mid i = 1, \ldots\ldots, N\}$ 学习得到的。为此，我们设计了一种优化方案，以便学习多模态位姿分布，这正是模糊场景所需要的。

1) 目标项：**预测误差**用于衡量预测位姿 $\hat{}y \in$SE(3) 与其真实值y之间的差异。我们将预测误差定义为平移误差项与旋转误差项的加权和，前者定义在 $\mathbb{R}^3$ 上，后者定义在 SO(3) 上。平移误差是预测位姿与真实位姿的平移分量 $\hat{}\mathbf{t}$，$\mathbf{t} \in \mathbb{R}^3$ 之间的欧几里得距离。对于旋转误差，我们选择旋转分量 $\hat{}\mathbf{R}$，$\mathbf{R} \in$ 之间的弦距离，因为与测地距离等方法相比，它在基于梯度的优化中具有更理想的梯度。因此，预测误差定义为

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

其中， λ_t 和 λ_r 是可调常数，用于平衡这两项的权重。

Kullback–Leibler 散度 $D_{\text{KL}}(q(\mathbf{z} \mid \mathbf{x}) \parallel p(\mathbf{z}))$用于衡量推断得到的潜变量后验分布$q(\mathbf{z} \mid \mathbf{x})$与定义在潜变量 $p(\mathbf{z})$上的先验分布之间的差异。这是变分原理的重要组成部分；它与预测误差共同构成在变分方法中优化的证据下界（ELBO）。按照惯例，为了便于计算 KL 散度，我们假设先验分布为标准高斯先验 $p(\mathbf{z}) = \mathcal{N}(\mathbf{z};\ \mathbf{0}, \mathbf{I})$。

2) 证据下界（*ELBO*）：在变分方法中，通常最大化的 ELBO 优化目标由负 KL 散度和预测的期望对数似然 $\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p_\phi(y \mid \mathbf{z})]$ 组合而成。后者的期望通常通过对 $q(\mathbf{z} \mid \mathbf{x})$ 进行蒙特卡洛模拟来计算。采用我们定义的姿态预测误差后，变分优化目标可写为

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

其中， $\mathcal{Z}_i = \{\mathbf{z}_j \sim q(\mathbf{z} \mid \mathbf{x}_i) \mid j = 1, \ldots, M\}$ 是蒙特卡洛样本集， $|\mathcal{Z}_i|$ 是其基数。

我们认为，在我们的设定中，最小化这一优化目标，尤其是期望预测误差，会适得其反；在模糊场景中，相机姿态后验分布 $p(y \mid \mathbf{z})$ 可能呈多峰分布。在这类情形下，两幅视觉上相似的图像 $\mathbf{x}_i$ 和 $\mathbf{x}j$ $(i \neq j)$ 会被编码为相似的潜在后验分布 $p(\mathbf{z} \mid \mathbf{x}_i)$ 和 $p(\mathbf{z} \mid \mathbf{x}j)$。然而，这些图像可能是在场景中的两个不同位姿 y_i 和 y_j 下拍摄的，此时相机位姿的真实后验分布 $p(y \mid \mathbf{x}_i)$ 和 $p(y \mid \mathbf{x}j)$ 均为双峰分布。最小化期望预测误差会产生折中解，即两个真实众数之间的单峰推断分布。我们提出对期望误差项进行修改，以解决这一问题。

3) 赢家通吃优化：我们提出，将计算出的平均预测误差限制在一组蒙特卡罗样本 $\mathcal{Z}_i \subseteq \mathcal{Z}_i$中，这些样本经映射$f_\phi(\cdot)$得到的图像与真实众数 y_i之间的距离小于某个阈值 δ，即 $\hat{}\ \mathcal{Z}_i = \{\mathbf{z}_j \in \mathcal{Z}_i \mid d_{\text{pose}}(f_\phi(\mathbf{z}j)\ y_i) < \delta\}$。这样可确保在优化过程中，位姿样本能够聚集在各个众数附近，而不受其他众数的影响。然而，真实后验分布未知，且不同众数的形状可能各异，因此 δ 的选择并非易事。此外，参数 θ和 ϕ 的随机初始化并不能保证在优化开始时，众数附近一定存在位于任意 δ 距离范围内的位姿样本。因此，需要在每次迭代中为每个众数自适应地选择 δ。

在每次迭代中，对于真实姿态 y_i，我们将$\delta_{i,\alpha}$ 选为以 y_i 为中心、包含 $\mathcal{Z}_i$中比例为 α 的样本的最小球的半径。换言之，我们定义的自适应$\delta_{i,\alpha}$为

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

导致仅针对每个真实姿态 y_i对应的蒙特卡洛样本中最近的 α 部分进行预测误差最小化。

我们提出的优化目标为

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

其中 $\hat{\mathcal{Z}}_{i,\,\alpha} = \{\mathbf{z}_j \in \mathcal{Z}_i \mid d_{\mathrm{pose}}(f_\phi(\mathbf{z}j)\;\; y_i) < \delta_{i,\;\alpha}\}$。 α 和 β 为可调常数，后者是 KL 散度项的平衡权重。

从思想上说，这与用于学习混合模型的赢家通吃多假设优化方案类似，即针对每个标签 [16], [33]优化距离最近的混合成分。然而，我们提出的解决方案处于不同的设置中，因为我们用样本而非混合模型来表示后验分布。因此，我们将我们的方法称为赢家通吃，以体现这种相似性，同时反映出该方法用于优化样本集，而非单个混合成分。

IV. 实验

A. 实现细节

我们使用 PyTorch 库 [34] 实现了我们的方法。我们采用 ResNet-18 [35] 作为骨干网络，并通过 编码器提取 2048 维特征向量，随后接入线性层，预测用于潜在后验分布的 d 维 $\boldsymbol{\mu}$ 和 $\log \boldsymbol{\sigma}^2$ 向量。*PoseMap* 由一个全连接网络实现，该网络将输入向量依次经过维度变换 $d \to 128 (\to 128)^{\times n_{\text{layers}}} \to 3 + 6$，中间采用 ReLU 激活函数。最少隐藏层数 n_{layers} 取决于场景中目标姿态分布的复杂程度。在几乎所有测试场景中，仅使用 $n_{\text{layers}} = 3$ 就取得了良好的性能；除非另有说明，所有实验均采用这一设置。最后一层用于预测平移和旋转向量，其中平移向量经过 Sigmoid 激活函数，随后通过固定的仿射变换对预测结果进行平移和缩放，使其落在场景的度量范围内。

我们使用 Adam 优化器 [36] 进行网络训练，初始学习率为 1×10^{-4}，每隔 $n_{\text{lr-decay}}$ 个周期将学习率按 0.8 的指数衰减率降低一次，共衰减 10 次。遵循姿态回归领域的惯例，我们首先调整每张图像的大小，使其最短边为 256，然后随机裁剪出 224×224 个区域作为编码器的输入。我们还通过调整颜色/亮度和应用高斯模糊来增强数据，以适应图像之间的光照变化和运动模糊。除非另有说明，我们均设定 $\alpha = 0.20$、$\beta = 0.01$，使用 $d = 16$ 维潜在空间，并在所有实验中用 1000 个蒙特卡洛样本表示分布，因为我们发现这在我们的设置下能产生良好的预测结果。其他超参数列于表 I 中，其设置经过调整，以反映不同数据集的图像数量和度量尺度；这些数据集涵盖了从小型室内场景到大型室外场景。请注意，我们在没有进行大规模超参数搜索的情况下确定了这些设置，并且

表 I 训练所用的超参数

数据集	λ_t	λ_r	批量大小	训练轮数	$n_{\text{lr-decay}}$
7-Scenes [24]	5	10	64	100	10
Cambridge Land. [10]	5	100	64	500	50
模糊重定位 [16]	5	2	4	500	50
天花板	5	2	4	2000	50
合成	5	2	4	500	50

通过全面搜索最优超参数，有望进一步提升性能。

B. 数据集和指标

我们在模糊重定位数据集 [16] 上评估我们的方法。该数据集是一个现有基准，包含来自模糊环境的真实图像序列。数据集中的每个场景都有单独的训练和测试图像序列，这些序列分别沿各自独特但总体视角相似的相机轨迹采集。我们发现，尽管人眼看来存在明显的歧义性，但该数据集中很大一部分帧包含独特的识别特征，表达能力强的特征检测器可以据此推断位姿。因此，大量帧的预测后验分布呈单峰，这妨碍了对方法形成多模态分布能力的评估。为此，我们在该数据集的基础上，新增录制了一段由机器制造面板构成的天花板真实序列，以呈现视觉模糊性极强的情况。我们使用标定后的激光雷达-惯性测量单元-相机装置录制训练和测试序列，并使用 MILIOM [37]获取真实相机位姿。此外，我们还渲染了来自 3D Warehouse[3],的两个合成场景的图像序列，这些场景在设计上具有对称性，并用它们在受控条件下研究我们的方法。

我们使用召回率作为评估模糊场景中位姿分布的指标。对于一张查询图像，我们从其后验分布中抽取样本；如果至少有 γ 比例的样本与真实姿态的距离在指定范围内，则将其视为真正例（否则视为假负例）。我们认为，对于众数间距较大且概率相等的分布，将 γ 设为与众数数量成反比，可以估计该分布在真实姿态周围是否具有足够的概率密度。除天花板场景外，我们对所有测试场景均报告采用 $\gamma = 0.1$ 时的召回率；对于天花板场景，则使用 $\gamma = 0.05$。

为验证我们的方法作为通用位姿回归器在无歧义场景中的表现，我们在视觉定位基准 7-Scenes [24]和 Cambridge Landmarks [10]上对其进行评估。与位姿回归研究中常见的做法一致，我们使用中位误差评估这些数据集上的表现。对于每个预测分布，我们从蒙特卡洛样本中计算点预测，分别采用平移的算术均值和方向的弦均值 L_2 [38]。我们报告该估计值与真实姿态之间的误差在每个场景上的中位数。

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

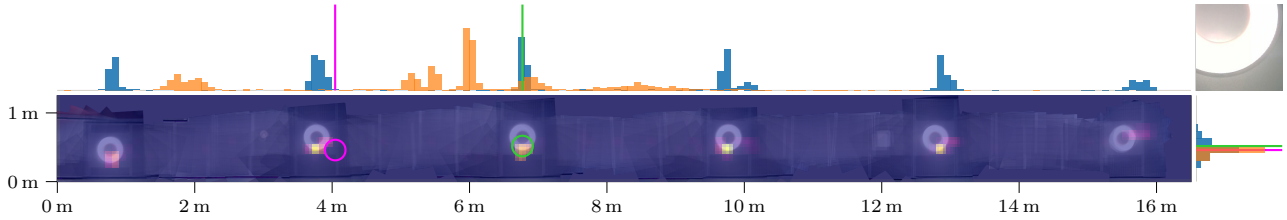


图 3. 我们的方法（）、Bingham MDN [16]（）、MapMet 的预测 [13]（）以及真实值（）对天花板场景中一张查询图像（右上）的 x 轴（左上）和 y 轴（右下）上的边缘后验分布预测。热力图显示了我们的方法预测的二维直方图，并将其叠加在该场景的拼接图像上。请注意，我们的方法成功捕捉到了分布的全部六个众数，而 MapNet 仅在错误位置预测了单个估计值，Bingham MDN 方法则在视觉上差异明显的位置赋予了较大的概率。

表 II 模糊场景中的测量召回率（越高越好）

场景	阈值	PN [10]	MN [13]	BMDN [16]	Abl.	Ours
蓝色椅子	0.1米/10°	0.08	0.05	0.41	0.32	**0.45**
	0.2米/15°	0.40	0.33	0.83	0.89	**0.99**
	0.3米/20°	0.56	0.46	0.89	0.97	**1.00**
会议桌	0.1米/10°	0.00	0.00	**0.09**	0.03	0.06
	0.2米/15°	0.02	0.03	0.27	0.24	**0.35**
	0.3米/20°	0.02	0.07	0.33	0.34	**0.43**
楼梯间	0.1米/10°	0.00	0.07	0.24	0.12	**0.30**
	0.2米/15°	0.01	0.17	0.48	0.44	**0.62**
	0.3米/20°	0.01	0.29	0.69	0.63	**0.72**
楼梯间外部	0.1米/10°	0.00	0.01	0.11	0.03	**0.12**
	0.2米/15°	0.00	0.03	0.43	0.24	**0.53**
	0.3米/20°	0.01	0.07	0.60	0.44	**0.71**
研讨室	0.1米/10°	0.00	0.09	0.38	0.17	**0.43**
	0.2米/15°	0.02	0.37	0.79	0.53	**0.90**
	0.3米/20°	0.10	0.53	0.91	0.80	**0.97**
天花板†	0.1米/10°	0.00	0.02	0.08	0.00	**0.09**
	0.2米/15°	0.03	0.05	0.19	0.02	**0.31**
	0.3米/20°	0.06	0.09	0.30	0。04	0。**44**

† 我们使用 $\alpha = 0.05$ 和 $n =$ 层来训练独立记录的天花板场景，这反映了场景中存在更丰富的模糊性。

表 III 无歧义场景中的中位误差（米/ °）（越低越好）

场景	PN [10]	MN† [13]	BPN [28]	BMDN [16]	Ours
国际象棋	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
头部	0.29/**12.0**	0.18/13.3	0.31/**12.0**	**0.13**/13.4	0.17/14.5
办公室	0.48/7.68	**0.17/5.15**	0.48/8.04	0.19/9.73	0.24/9.30
南瓜	0.47/8.42	0.22/**4.02**	0.61/7.08	**0.20**/9.40	0.30/8.33
厨房	0.59/8.64	0.23/**4.93**	0.58/7.54	**0.19**/10.9	0.26/10.2
楼梯	0.47/13.8	**0.30/12.1**	0.48/13.1	0.34/14.1	0.47/15.5
大学	1.92/5.40	**1.07/1.89**	1.74/4.06	1.51/2.14	1.65/2.88
街道	3.67/6.50	–	**2.96/6.00**	16.3/25.2	17.2/23.8
医院	2.31/5.38	**1.94/3.91**	2.57/5.14	2.25/3.93	2.06/4.33
立面	1.46/8.08	1.49/**4.22**	1.25/7.54	3.52/5.41	**1.02**/6.03
教堂	2.65/8.48	2.00/**4.53**	2.11/8.38	2.16/5.99	**1.80**/5.90

†摘自 Sattler 等人 [23] 的 Cambridge Landmarks 数据集上的 MapNet 结果。

C. 基准数据集上的评估

我们在表 II 中报告了模糊场景上的结果。可以看到，我们的方法优于 Bingham MDN [16]；后者是与我们的方法最接近的、旨在解决模糊场景定位问题并预测位姿分布的方法。我们考察了该方法的两种设置，其混合模型分别包含 10 个和 50 个分量，并根据从中抽取的样本评估指标。由于包含 10 个分量的设置始终表现更好，我们在表中将其结果作为代表（标记为 BMDN）。图 3 展示了一个示例：给定来自天花板场景的查询图像时，模型预测的后验分布。可以看出，我们的方法预测的后验分布更好地捕捉了场景的模糊结构。我们还评估了 PoseNet[10]及其贝叶斯变体 [28]，以及 MapNet [13]。然而，我们发现这些单点估计方法在模糊场景上的表现无法与之媲美。令人意外的是，原始 PoseNet 的表现相对优于贝叶斯 PoseNet，因此我们将其结果作为代表（标记为 PN），并与 MapNet（标记为 MN）的结果一同列出。

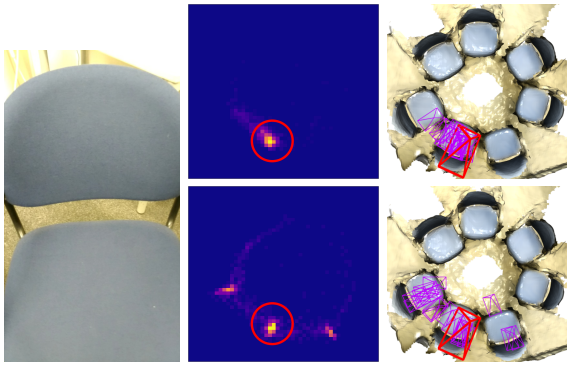
为了探究我们方法性能提升是否源于采用变分推断的新颖形式化，我们进行了消融实验，将流程修改为针对输入图像生成单个位姿。

我们从目标函数中移除 KL 散度项，修改编码器 以预测单个点，并将编码器的预测结果输入*PoseMap*，从而得到单个位姿预测。在其他条件相同的情况下，我们评估了这一消融变体；原则上，它与 PoseNet 非常相似。从表 II 中可以看出，该变体标记为消融，虽然凭借更新的特征提取网络，表现优于 PoseNet，但仍不及未经消融的变体，这验证了我们所提出形式化方法的价值。

为求完整，我们在表 III 中报告了无歧义的 7-Scenes 和 Cambridge Landmarks 数据集上的结果。我们将 PoseNet 和 MapNet 作为单一位姿回归器基线，并将贝叶斯 PoseNet 和 Bingham MDN 作为原则上能够预测多模态分布的方法纳入比较。虽然我们的方法并非表现最佳，但与顶尖方法相差不远。此实验仅用于检验我们的方法在最简流程中的性能，不包含任何旨在提升无歧义场景中准确率的特定机制。如表 II 所示，在无歧义场景中表现较好的方法，在模糊场景中的表现较差，而这正是我们的方法旨在解决的问题。未来一个值得探索的方向是，将我们提出的、旨在处理模糊场景的公式与提升无歧义场景中姿态回归性能的技术结合起来。

D. 深入探究

图4（上）展示了我们的方法对 Ambiguous Relocalization 数据集中的一个查询图像示例所预测的分布。尽管这个由外观完全相同的椅子组成的场景在人眼看来或许存在歧义，但我们可以看到



(a) 查询图像 (b) 位置后验 (c) 样本

图4。我们的方法在完整容量（上）以及通过 L_2 权重衰减 $\lambda = 0.1$ 进行受限学习模式（下）下预测的后验分布。后者能够捕捉多个众数，而前者只预测一个位于正确位姿处的众数。

预测的后验分布识别出了正确的位姿，并将概率密度集中在其周围。我们推测，表达能力足够强的 编码器能够凭借最细微的细节，区分现实场景中拍摄的、看似模棱两可的图像，例如本例中椅子背景的差异。然而，表达能力较弱的数据 编码器无法学习所有细节，因此可能会受到歧义的影响。我们通过在训练期间对 编码器权重的 L_2 范数施加惩罚项来检验这一假设。从图4（下）可以看出，这种设置使预测的后验分布为能看到另外两把椅子的位姿分配了概率。我们认为，当训练数据与运行条件之间存在领域差距时，模型应当适当牺牲预测置信度，以提升泛化能力，而有意施加学习约束可以实现这一点。我们将此类学习约束的研究留待未来开展。

我们在两个合成场景中研究了赢家通吃优化方案中 α 的影响：一个场景中，相机围绕一张有四条桌腿的圆桌移动，导致图像的位姿分布出现四个众数；另一个场景中，相机围绕一张长方形餐桌移动，导致分布呈双峰。我们报告了10次训练运行的统计结果：使用不同的 α 值进行训练后，在训练结束时评估$0.1\mathrm{m}/10^\circ$ 召回率，结果见图6。可以看出，在这些场景中，当 α 取值大于零且小于众数数量 $1/\#$时，召回率最高。图5展示了 α取三个不同值时预测的相机位置后验分布。正如第III-C.2节所述， α过大会导致后验分布受损，而 α 过小则会在训练数据覆盖范围内预测出近乎均匀的密度。我们推测，赢家通吃优化要收敛并捕获分布中的所有众数， α 必须小于众数数量 $1/\#$；同时， α 还必须充分大于零，才能克服小批量优化带来的噪声。在这一 α 取值范围内，训练速度与学习到的分布质量之间存在权衡： α 越小，每次迭代优化的样本就越少，因此训练速度越慢，但较不容易

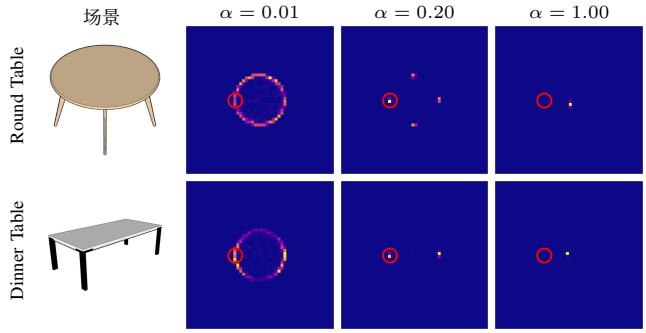
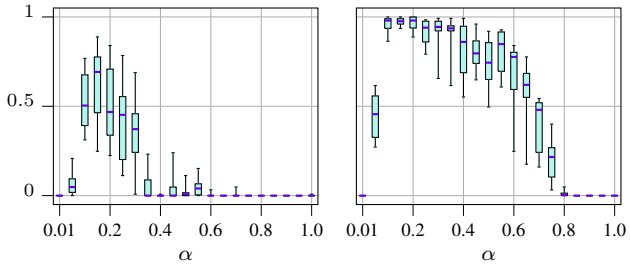


图5。合成场景中针对 α 的不同取值，在 xy平面上预测的位置后验。圆桌和餐桌的真实分布分别有四个和两个众数。当 $0 \ll \alpha < 1/\#$ 个众数时，优化能够成功收敛并预测出正确的众数。例如， $\alpha = 0.20$ 对两个场景都能给出良好的预测。



(a) 圆桌 (b) 餐桌

图6. 在合成场景的赢家通吃优化中，针对不同的 α 选择，阈值为 $0.1\mathrm{m}/10^\circ$ 时的测量召回率。对于每种 α选择，统计 10 次训练运行的结果：箱体从召回率的下四分位数延伸至上四分位数，紫色线表示召回率中位数，须线则延伸至最小值和最大值。

受到蒙特卡洛采样所引入噪声的影响。我们将寻找最优 α 留待未来工作研究。

E. 运行时间评估

我们在配备 Intel Core i7-8700K CPU 和 NVIDIA GeForce GTX 1080 Ti GPU 的台式电脑上，测量一个查询图像通过我们的流程进行 1000 个蒙特卡洛样本前向传播所需的时间。每项测量重复 100 次，结果显示，前向传播在 CPU 上平均耗时 $14.88 \pm 0.75\,\mathrm{ms}$，在 GPU 上平均耗时 $2.26 \pm 0.10\mathrm{ms}$；也就是说，我们的流程可以实时运行。

V. 结论

在这项工作中，我们研究了模糊场景中的视觉定位任务。我们提出了一种结合变分推断的相机位姿回归新形式化方法，使模型能够学习并从给定图像所对应的所有相机位姿分布中采样。具体而言，首先对图像进行编码，预测场景中视觉特征潜在空间上的后验分布。从该分布中抽取样本，并将其输入学习得到的映射，即可生成一组位姿样本，用以表示相机位姿的后验分布。我们展示了，与现有方法相比，我们的形式化方法在模糊场景定位任务中表现更优，并提出了未来研究方向，以进一步考察所提出的方法。

参考文献

[1] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, 和 T. Pajdla, “通过视图合成实现全天候地点识别”, 载于《IEEE计算机视觉与模式识别会议论文集》, 2015年, 第1808–1817页。[2] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, 和 J. Sivic, “NetVLAD：用于弱监督地点识别的卷积神经网络架构”, 载于《IEEE计算机视觉与模式识别会议论文集》, 2016年, 第5297–5307页。 [3] Z. Chen, A. Jacobson, N. S¨underhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, 和 M. Milford, “用于视觉地点识别的大规模深度学习特征”, 载于《IEEE机器人与自动化国际会议论文集》, 2017年, 第3223–3230页。 [4] S. Hausler, S. Garg, M. Xu, M. Milford, 和 T. Fischer, “Patch-NetVLAD：用于地点识别的局部-全局描述符多尺度融合”, 载于《IEEE/CVF计算机视觉与模式识别会议论文集》, 2021年, 第14 141–14 152页。 [5] Y. Li, N. Snavely, D. Huttenlocher, 和 P. Fua, “利用三维点云进行全球位姿估计”, 载于《欧洲计算机视觉会议论文集》。施普林格, 2012年, 第15–29页。 [6] T. Sattler, B. Leibe, 和 L. Kobbelt, “通过主动对应搜索改进大规模基于图像的定位”, 《IEEE模式分析与机器智能汇刊》, 第39卷, 第9期, 第1744–1756页, 2016年。[7] L. Liu, H. Li, 和 Y. Dai, “在大规模三维地图中实现高效的全局二维-三维匹配以进行相机定位”, 载于《国际计算机视觉会议论文集》, 2017年, 第2372–2381页。 [8] P.-E. Sarlin, A. Unagar, M. Larsson, H. Germain, C. Toft, V. Larsson, M. Pollefeys, V. Lepetit, L. Hammarstrand, F. Kahl 等, “回归特征：从像素到位姿学习鲁棒的相机定位”, 载于《IEEE/CVF计算机视觉与模式识别会议论文集》, 2021年, 第3247–3257页。 [9] T. Sattler, B. Leibe, 和 L. Kobbelt, “通过主动对应搜索改进基于图像的定位”, 载于《欧洲计算机视觉会议论文集》。施普林格, 2012年, 第752–765页。 [10] A. Kendall, M. Grimes, 和 R. Cipolla, “PoseNet：用于实时六自由度相机重定位的卷积网络”, 载于《IEEE计算机视觉与模式识别会议论文集》, 2015年, 第2938–2946页。 [11] R. Clark, S. Wang, A. Markham, N. Trigoni, 和 H. Wen, “VidLoc：用于六自由度视频片段重定位的深度时空模型”, 载于《IEEE计算机视觉与模式识别会议论文集》, 2017年, 第6856–6864页。 [12] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, 和 D. Cremers, “利用长短期记忆网络进行结构化特征相关性的基于图像的定位”, 载于《国际计算机视觉会议论文集》, 2017年, 第627–637页。 [13] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, 和 J. Kautz, “面向相机定位的地图几何感知学习”, 载于《IEEE/CVF计算机视觉与模式识别会议论文集》, 2018年, 第2616–2625页。 [14] S. Chen, Z. Wang, 和 V. Prisacariu, “Direct-PoseNet：结合光度一致性的绝对姿态回归”, 载于《国际三维视觉会议论文集》。IEEE, 2021年, 第1175–1185页。 [15] A. Moreau, N. Piasco, D. Tsishkou, B. Stanciulescu, 和 A. de La Fortelle, “LENS：通过神经辐射场合成增强定位”, 载于《机器人学习会议》。PMLR, 2022年, 第1347–1356页。 [16] H. Deng, M. Bui, N. Navab, L. Guibas, S. Ilic, 和 T. Birdal, “深度宾厄姆网络：处理姿态估计中的不确定性与歧义性”, 《国际计算机视觉期刊》, 第1–28页, 2022年。 [17] D. Fox, S. Thrun, W. Burgard, 和 F. Dellaert, “ 用于移动机器人定位的粒子滤波器”, 载于《实践中的序贯蒙特卡罗方法》。施普林格, 2001年, 第401–428页。 [18] P. Jensfelt 和 S. Kristensen, “利用多假设跟踪实现移动机器人的主动全局定位”, 《IEEE机器人与自动化汇刊》, 第17卷, 第5期, 第748–760页, 2001年。 [19] D. Fox, “KLD采样：自适应粒子滤波器”, 载于《神经信息处理系统进展》, T. Dietterich, S. Becker, 和 Z. Ghahramani 编, 第14卷。麻省理工学院出版社, 2001年。 [20] K. A. Murphy, C. Esteves, V. Jampani, S. Ramalingam, 和 A. Maka-dia, “Implicit-PDF：旋转流形上概率分布的非参数表示”, 载于《国际机器学习会议论文集》, 2021年, 第7882–7893页。

[21] D. P. Kingma 和 M. Welling, “自编码变分贝叶斯”, 载于国际学习表征会议论文集, 2014年。[22] D. J. Rezende、S. Mohamed 和 D. Wierstra, “深度生成模型中的随机反向传播与近似推断”, 载于国际机器学习会议论文集。PMLR, 2014年, 第1278–1286页。[23] T. Sattler、Q. Zhou、M. Pollefeys 和 L. Leal-Taixe, “理解基于CNN的绝对相机位姿回归的局限性”, 载于*IEEE/CVF*计算机视觉与模式识别会议论文集, 2019年, 第3302–3312页。[24] J. Shotton、B. Glocker、C. Zach、S. Izadi、A. Criminisi 和 A. Fitzgibbon, “用于RGB-D图像相机重定位的场景坐标回归森林”, 载于*IEEE*计算机视觉与模式识别会议论文集, 2013年, 第2930–2937页。[25] W. Kabsch, “用于确定两组向量之间最佳旋转的解”, 《晶体学报*A*辑：晶体物理、衍射、理论与一般晶体学》, 第32卷, 923, 1976年5, 922第。–页。[26] T. Naseer 和 W. Burgard, “用于室外环境中单目相机六自由度全局定位的深度回归”, 载于*IEEE/RSJ*智能机器人与系统国际会议论文集, 2017年, 第1525–1530页。[27] T. Ng、A. Lopez-Rodriguez、V. Balntas 和 K. Mikolajczyk, “重新评估CNN方法在相机位姿回归中的局限性”, *arXiv*预印本 *arXiv:2108.07260*, 2021年。[28] A. Kendall 和 R. Cipolla, “深度学习中相机重定位的不确定性建模”, 载于*IEEE*国际机器人与自动化会议论文集。IEEE, 2016年, 第4762–4769页。[29] ——, “深度学习相机位姿回归的几何损失函数”, 载于*IEEE*计算机视觉与模式识别会议论文集, 2017年, 第5974–5983页。[30] A. Moreau、N. Piasco、D. Tsishkou、B. Stanciulescu 和 A. de La Fortelle, “CoordiNet：用于可靠车辆定位的不确定性感知位姿回归器”, 载于*IEEE/CVF*计算机视觉与模式识别会议论文集, 2022年, 第2229–2238页。[31] Y. Xie、T. Takikawa、S. Saito、O. Litany、S. Yan、N. Khan、F. Tombari、J. Tompkin、V. Sitzmann 和 S. Sridhar, “视觉计算及其他领域中的神经场”, 《计算机图形学论坛》, 第41卷, 第2期, 第641–676页, 2022年。[32] Y. Zhou、C. Barnes、J. Lu、J. Yang 和 H. Li, “神经网络中旋转表示的连续性”, 载于*IEEE/CVF*计算机视觉与模式识别会议论文集, 2019年, 第5745–5753页。[33] O. Makansi、E. Ilg、O. Cicek 和 T. Brox, “克服混合密度网络的局限性：用于多模态未来预测的采样与拟合框架”, 载于*IEEE/CVF*计算机视觉与模式识别会议论文集, 2019年, 第7144–7153页。[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 和 S. Chintala, “PyTorch：一款命令式风格、高性能的深度学习库”, 载于神经信息处理系统进展, H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alch´e-Buc、E. Fox 和 R. Garnett 编。Curran Associates, Inc., 2019年, 第8024–8035页。[35] K. He、X. Zhang、S. Ren 和 J. Sun, “用于图像识别的深度残差学习”, 载于*IEEE*计算机视觉与模式识别会议论文集, 2016年, 第770–778页。[36] D. P. Kingma 和 J. Ba, “Adam：一种随机优化方法”, 载于国际学习表征会议论文集, 2015年。[37] T.-M. Nguyen、S. Yuan、M. Cao、L. Yang、T. H. Nguyen 和 L. Xie, “Miliom：紧耦合多输入激光雷达-惯性里程计与建图”, 《*IEEE*机器人与自动化快报》, 第6卷, 第3期, 第5573–5580页, 2021年。[38] R. Hartley、J. Trumpf、Y. Dai 和 H. Li, “旋转平均”, 《国际计算机视觉期刊》, 第103卷, 第3期, 第267–305页, 2013年。