This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Probabilistic Perspective-n-lines for Indoor Camera Pose Estimation

Xiaowei Chen
University of Central Arkansas
Conway, AR 72035, USA
xchen16@uca.edu

Guoliang Fan
Oklahoma State University
Stillwater, OK 74078, USA
guoliang.fan@okstate.edu

Abstract

Indoor localization from a single RGB image via Perspective-n-Lines (PnL) remains a fundamental yet challenging problem in computer vision. To address this, the PnL-IOC algorithm was introduced, demonstrating remarkable performance by jointly estimating 3D correspondences of image outer corners (IOCs)—the intersection points between image borders and room layout boundaries—while optimizing camera pose within an iterative Gauss-Newton framework. However, existing learning-based indoor layout estimation methods often struggle to achieve high accuracy, resulting in unreliable line correspondences, which significantly limits the effectiveness of PnL-IOC. To overcome this limitation, we propose InPro-PnL, a probabilistic PnL layer for indoor camera pose estimation. Our approach introduces a novel framework that models the pose distribution on the SE(3) manifold, effectively extending the categorical Softmax function into the continuous domain. This is achieved by integrating a probabilistic PnL layer, where denoised 2D-3D correspondences serve as intermediate variables and are optimized by minimizing the KL divergence between the predicted and target pose distributions. Experimental results demonstrate that our method outperforms the geometric PnL-IOC approach and holds significant potential for further enhancing room layout estimation accuracy.

1. Introduction

Room-level indoor localization is essential for egocentric vision-based assistive tools in home environments, enabling location-aware assistance and supporting indoor navigation[35, 44, 46, 47]. Given that most indoor structures conform to the Manhattan world assumption[19], indoor scene images can be effectively represented by structured room layouts, which are characterized by a finite set of boundary lines and junctions, commonly referred to as image inner corners (IICs) [37]. Recent studies [15, 16] introduced the PnL-IOC method for indoor localization, leveraging room layouts while incorporating image outer corners (IOCs)—the intersection points between room layout boundaries and image borders—to introduce additional auxiliary lines for efficient camera pose estimation. This approach relies on n correspondences between 3D room layout lines and their 2D projections to achieve accurate pose estimation. Building on this work, we formulate camera pose estimation from room layouts as a Perspective-n-Lines (PnL) problem [56], integrating a priori constraints, such as room dimensions, to improve pose estimation accuracy. Furthermore, since most indoor environments adhere to the Manhattan world assumption, this assumption naturally facilitates the definition of 3D line-based representations for room layouts.

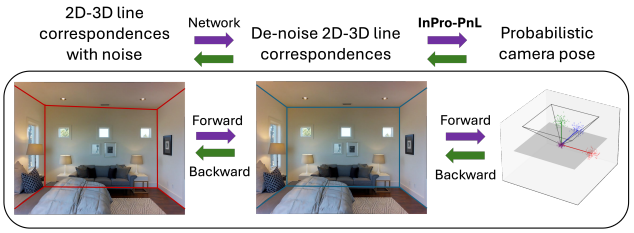


Figure 1. InPro-PnL is a solution to 2D-3D line correspondence learning.

The recent PnL-IOC method has demonstrated high accuracy and stability in camera pose estimation. However, the 2D/3D layout line correspondences used in this approach are derived from ground truth data with varying levels of noise. Obtaining accurate ground truth can be challenging, despite the rapid advancements in indoor layout estimation, particularly with the emergence of deep learning-based methods over the past decade, which have shown significant promise and potential [36, 41, 57]. To address these limitations, we propose an indoor probabilistic PnL (InPro-PnL) approach that leverages learning-based denoising of 2D-3D line correspondences (Figure 1), inspired by recent state-of-the-art methods [14, 32, 62]. The core concept is simple: while deterministic pose estimation is non-differentiable, the probability density of pose, similar to cat-

egorical classification scores, is inherently differentiable. To leverage this property, we represent the output of PnL as a probabilistic distribution, parameterized by de-noised 2D-3D line correspondences. During training, we use the Kullback-Leibler (KL) divergence between the predicted and target pose distributions as the loss function, which is efficiently computed through Monte Carlo pose sampling.

The rest of this paper is structured as follows. In Section 2, the related work is briefly reviewed. An overview of the proposed method is presented in Section 3. In Section 4, we discuss the network we use to test the InPro-PnL layer. Experimental results are shown in Section 5. Finally, we draw conclusions in Section 6.

2. Related work

The related work is briefly reviewed from three perspectives: *room layout estimation*, the *PnL* problems, *end-to-end correspondence learning*, and *Probabilistic Deep Learning*.

Room layout estimation has been an active research topic for decades. Early approaches primarily relied on geometry-based methods, leveraging vanishing point estimation to infer room structures [27, 42, 43, 48, 53]. With the advancement of deep learning, a set of well-defined layouts has enabled robust and accurate representation of a wide range of complex indoor scenes [28, 37, 41, 57, 59]. Additionally, the release of high-quality datasets [20, 30, 60] has further propelled the development of deep learning-based methods, enhancing room layout estimation and supporting camera pose estimation for location-aware vision tasks. Building on room layout estimation, indoor camera pose estimation can be formulated as a Perspective-n-Line (PnL) problem.

In the Perspective-n-Line (PnL) problem, at least three 2D/3D line correspondences are required since a 3D camera pose has six degrees of freedom (DoF), with each line correspondence providing two constraints [25]. Most existing PnL approaches address cases where $n > 3$ and can be broadly classified into iterative and non-iterative methods. Early iterative methods [17, 22, 34] are computationally intensive, sensitive to initial conditions, and prone to convergence towards local minima [54]. Recent non-iterative approaches have introduced linear formulations [1, 45, 50], which, although efficient, are sensitive to noise and perform poorly with small line sets ($n < 6$). Some non-iterative methods [2, 40] specifically target small sets but often compromise stability due to inherent linearization errors. To mitigate these limitations, a robust, non-iterative $O(n)$ solution, known as Robust PnL (RPnL), was proposed in [58] for cases with $n \geq 4$. Building upon RPnL, the Accurate Subset-based PnL (ASPnL) method [54] improved accuracy specifically for small line sets. However, ASPnL faces challenges when applied to scenarios with exactly three orthogonal lines intersecting at a single junction point. This limitation was addressed by the SRPnL method [56], which extends ASPnL to handle such cases effectively, yielding high accuracy on small line sets. Nevertheless, SRPnL encounters difficulties with larger line sets ($n \geq 8$) and under conditions of significant noise. Further enhancing robustness, the uncertainty-aware PnL method (DLSLU*) introduced by [52] incorporates uncertainty modeling in both 3D coordinates and their 2D projections to optimize existing PnL solutions. However, this method requires at least three known point correspondences to construct the line correspondences. In contrast, the PnL-IOC method proposed by [16] is tailored explicitly for indoor camera pose estimation, leveraging prior knowledge such as room dimensions. This method can solve the PnL problem even when provided with only one point correspondence. Nevertheless, it should be noted that the effectiveness of all the methods described above critically depends on the accuracy of the input information. In practice, however, such information is typically derived from deep learning or automatic detection methods, which do not always guarantee high precision.

To mitigate noise introduced by earlier processing stages, end-to-end correspondence learning approaches have been proposed, enabling seamless gradient backpropagation from pose estimates to intermediate representations. This paradigm has been fully exploited in recent object pose estimation studies. Specifically, by differentiating the Perspective-n-Point (PnP) operation, Brachmann and Rother [5] developed a dense correspondence network where 3D point coordinates are rendered learnable. Concurrently, BPnP [12] introduced an innovative prediction of 2D keypoint locations, whereas BlindPnP [10] proposed learning the corresponding weight matrix for unordered 2D-3D point pairs. Extending beyond traditional point correspondences, RePOSE [29] pioneered a feature-metric correspondence network, employing a similar end-to-end training strategy. Each of these methods utilizes a surrogate regularization loss to facilitate convergence, as the deterministic nature of pose estimation – being inherently non-differentiable – would otherwise hinder stable and reliable training.

Probabilistic approaches effectively handle the inherent uncertainties within models and data, distinguished as epistemic and aleatoric uncertainties, respectively [31]. Aleatoric uncertainty characterizes prediction variability by employing learnable probabilistic distributions. In classification tasks, discrete categorical distributions implemented via the Softmax function have become the prevalent method, serving as a smooth approximation to the one-hot arg-max operation. This concept has inspired innovative methods such as DSAC [8], a smooth variant of RANSAC operating over a finite hypothesis set. For quantifying continuous variables [13, 23, 26, 31, 33, 55], tractable paramet-

ric distributions—particularly Gaussian (normal) distributions—are commonly adopted. To better capture complex ambiguities [4, 7, 38], mixture models have been employed, notably for challenging problems such as 6DoF pose estimation [9]. Recently, [14] proposed an innovative method that backpropagates gradients through a continuous probability distribution derived from a nested optimization layer (the Perspective-n-Point, PnP, layer). This strategy acts as a continuous analogue to the Softmax function, significantly enhancing the model's precision in navigating continuous spaces. In this paper, we extend the concept by backpropagating a continuous distribution derived specifically from the Perspective-n-Line (PnL) layer, further advancing the model's capacity to manage complexities associated with continuous pose estimation.

3. Probabilistic Perspective-n-lines

3.1. Overview

Given a noisy set of line correspondences, our goal is to denoise this set, denoted as $L = \{\mathbf{l}_i^{3D}, \mathbf{l}_i^{2D}, \mathbf{w}_i^{3D} | i = 1...N\}$ of N corresponding lines, with 3D room layout line $\mathbf{l}_i^{3D} = (\mathbf{v}_i, \mathbf{P}_i)$, where $\mathbf{v}_i \in \mathbb{R}^3$ is the normalized vector giving the direction of the line and $P_i \in \mathbb{R}^3$ is any point on the line, 2D line $\mathbf{l}_i^{2D} \in \mathbb{R}^2$, which is the projection of $\mathbf{l}_i^{3D}$ on the image plane, with $\mathbf{p}_{is}$ and $\mathbf{p}_{ie}$ serving as the end points of $\mathbf{l}_i^{2D}$, and weights $\mathbf{w}_i^{3D}$ for the normal of a projection plane Π_i that is formed by the $\mathbf{l}_i^{3D}$, $\mathbf{l}_i^{2D}$ and the projection center O, and this normal can be defined as $\mathbf{n}_i^c$ and calculated from the cross product of $\mathbf{p}_{is}$ and $\mathbf{p}_{ie}$. From this set of line correspondences, a PnL problem can be formulated to estimate the indoor camera pose including the rotation $\mathbf{R}_w^c$, also short as $\mathbf{R}$, and translation $\mathbf{t}$, as Fig. 2 illustrated.

The core idea of a PnL layer is to find an optimal pose p, comprising a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$, that minimizes the cumulative squared orthogonal error:

$$\arg\min_{p} \frac{1}{2} \sum_{i=1}^{N} \Big(\left\| \mathbf{w}_i^{3D} (\mathbf{n}_i^c)^T (\mathbf{R}\mathbf{P}_i + \mathbf{t}) \right\|^2 + \left\| \mathbf{w}_i^{3D} (\mathbf{n}_i^c)^T \mathbf{R}\mathbf{v}_i \right\|^2 \Big). \quad (1)$$

Here, we let

$$f_i(p) = ||\mathbf{w}_i^{3D} (\mathbf{n}_i^c)^T (\mathbf{R}\mathbf{P}_i + \mathbf{t})|| + ||\mathbf{w}_i^{3D} (\mathbf{n}_i^c)^T \mathbf{R}\mathbf{v}_i||.$$

Eq. (1) represents a nonlinear least-squares problem that may yield non-unique solutions, leading to pose ambiguities[6, 21]. Previous studies [10, 12, 39, 49] typically perform backpropagation only through a locally optimal solution p^*, which is inherently unstable and non-differentiable. To facilitate a stable, differentiable alternative appropriate for end-to-end learning, one recent approach modeled the PnP output as a probabilistic distribution over poses, ensuring differentiability through a probability density function [14]. Inspired by this idea, we propose modeling the PnL output as a distribution over poses, where we define the cumulative error as the negative logarithm of the likelihood function $P(L|p)$, expressed as follows:

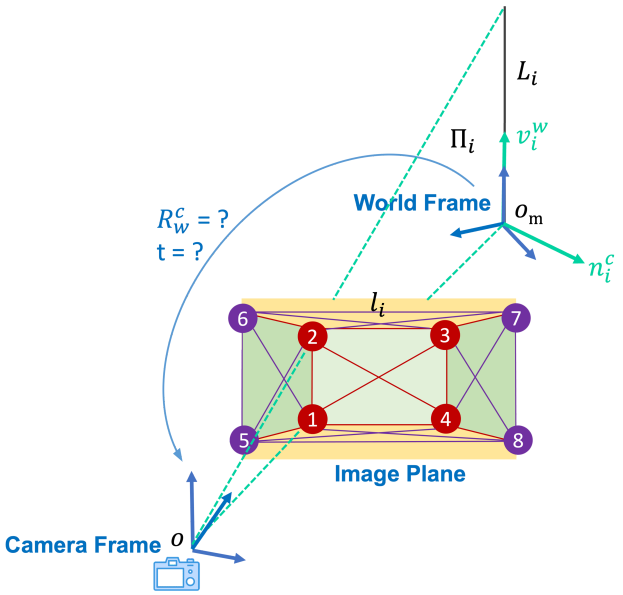


Figure 2. Illustration of the PnL problem.

$$P(L|p) = exp(-\frac{1}{2} \sum_{i=1}^{N} ||f_i(p)||^2). \quad (2)$$

By incorporating an additional prior pose distribution $P(p)$, we can derive the posterior pose $P(L|p)$ using Bayes' theorem. When an uninformative prior is used, the posterior density simplifies to the normalized likelihood:

$$P(p|L) = \frac{exp(-\frac{1}{2} \sum_{i=1}^{N} ||f_i(p)||^2)}{\int exp(-\frac{1}{2} \sum_{i=1}^{N} ||f_i(p)||^2)}. \quad (3)$$

Eq. (3) can be viewed as the continuous equivalent of the categorical Softmax.

3.1.1. KL Loss Function

During training, with a target pose distribution defined by the probability density $t(p)$, the training loss is minimized by reducing the KL divergence $D_{\mathrm{KL}}(t(p)||P(p|L))$. Intuitively, pose ambiguity is represented through the multiple modes of $P(p|L)$, and the loss function ensures convergence by suppressing incorrect modes. Omitting the constant, the KL divergence loss is expressed as:

$$L_{\mathrm{KL}} = -\int t(p) \log P(L|p) dp + \log \int P(L|p) dp. \quad (4)$$

It is proved by experiments that using a narrow (Dirac-like) target distribution centered at the ground truth p_{gt} is effective. This approach simplifies the loss function, as shown after substituting Eq. (2):

$$L_{\mathrm{KL}} = \frac{1}{2}\sum_{i=1}^{N}||f_i(p_{gt})||^2 + \log\int exp(-\frac{1}{2}\sum_{i=1}^{N}||f_i(p)||^2). \tag{5}$$

The first item in Eq. (5) represents the orthogonal error at target pose while the first item represents the orthogonal error at predicted pose. The only remaining issue is the integration in the second term, which is detailed in Section 3.2.

3.2. Monte Carlo Pose Loss

This section presents an efficient, GPU-compatible Monte Carlo method for integrating the proposed loss function, leveraging Adaptive Multiple Importance Sampling (AMIS) [14, 18]. Let $q(p)$ represent the probability density function of a proposal distribution designed to approximate the integrand $exp(-\frac{1}{2}\sum_{i=1}^{N}||f_i(p)||^2)$. With p_i as one of the K samples drawn from $q(p)$, the second term in Eq. (5), denoted as L_{prep}, can then be approximated as follows:

$$L_{pred} \approx \log\frac{1}{K}\sum_{j=1}^{K}\frac{exp(-\frac{1}{2}\sum_{i=1}^{N}||f_i(p)||^2)}{q(p_j)}. \tag{6}$$

We let

$$v_i = \frac{exp(-\frac{1}{2}\sum_{i=1}^{N}||f_i(p)||^2)}{q(p_j)},$$

which compactly denotes the importance weight at p_j. Eq.(6) describes the basic importance sampling method, where the selection of the proposal distribution $q(p)$ significantly influences the numerical stability. A more effective alternative, the AMIS algorithm, iteratively adjusts the proposal distribution to better align with the integrand.

In summary, AMIS leverages the importance weights from previous iterations to refine the proposal distribution. It then re-weights all past samples as if they were uniformly drawn from a cumulative mixture of all proposals used to date. The initial proposal is typically determined based on the mode and covariance of the predicted pose distribution.

Selection of the Proposal Distribution Proposal distributions for position and orientation are selected independently due to the non-Euclidean nature of orientation space. We use a 3-degree-of-freedom (DoF) multivariate t-distribution for position. For 3D orientation, represented by a unit quaternion, we adopt the angular central Gaussian distribution [14, 51].

3.3. Backpropagation

Generally, the partial derivative of the loss function defined in Eq. (5) is:

$$\frac{\partial L_{\mathrm{KL}}}{\partial(\cdot)} = \frac{\partial}{\partial(\cdot)}\frac{1}{2}\sum_{i=1}^{N}||f_i(p)||^2 - \mathop{\mathbb{E}}_{p\sim P(p|L)}\frac{\partial}{\partial(\cdot)}\frac{1}{2}\sum_{i=1}^{N}||f_i(p)||^2, \tag{7}$$

where the first term denotes the gradient of the orthogonal error evaluated at the target pose, and the second term approximates the gradient across the predicted pose distribution by backpropagating through each weighted sample obtained via Monte Carlo integration.

Balancing Uncertainty and Discrimination Examine the negative gradient relative to the associated weights w_i^{3D}:

$$-\frac{\partial L_{\mathrm{KL}}}{\partial \mathbf{w}_i^{3D}} = \mathbf{w}_i^{3D} \circ (-r_i^{\circ 2}(p_{gt}) + \mathop{\mathbb{E}}_{p\sim P(p|L)} r_i^{\circ 2}(p)), \tag{8}$$

where $r_i^{\circ 2}(p) = ||(\mathbf{n}_i^c)^T(\mathbf{R}\mathbf{P}_i + \mathbf{t})||^2 + ||(\mathbf{n}_i^c)^T\mathbf{R}\mathbf{v}_i||^2$ denotes the unweighted orthogonal errors, and $(\cdot)^{\circ 2}$ indicates element-wise squaring. The first term, $-r_i^{\circ 2}(p_{gt})$, carries a negative sign, suggesting that correspondences with larger orthogonal errors – reflecting higher uncertainty – should receive lower weights. The second term, $\mathop{\mathbb{E}}_{p\sim P(p|L)} r_i^{\circ 2}(p)$, captures the variance of orthogonal errors across the predicted pose distribution. The positive sign here implies that correspondences sensitive to pose variations should be assigned higher weights, as they significantly enhance discriminative capability. Consequently, the combined gradient effectively balances uncertainty management and discriminative enhancement.

3.4. Local Regularization of Derivatives

Although the KL divergence provides a suitable metric for measuring probabilistic distributions during training, inference still necessitates estimating the exact pose p^* by explicitly solving the PnL problem in Eq. (1). To achieve high precision, a practical choice is the recent developed PnL-IOC algorithm [16], which offers a robust linear solution. Specifically, during inference, the optimal pose is computed by applying the PnL-IOC method, followed by iterative refinement using the Jacobian matrix of the weighted orthogonal errors across all line correspondences. To facilitate convergence and improve robustness during training, a detached solution p^* is obtained first. Then, p^* is refined through the updating rule $p^* = p^* + \Delta p$. The increment Δp is determined by the Jacobian matrix of the weighted orthogonal errors of all lines. We therefore design the regularization loss as follows:

$$L_{reg} = l(p^* + \Delta p, p_{gt}), \tag{9}$$

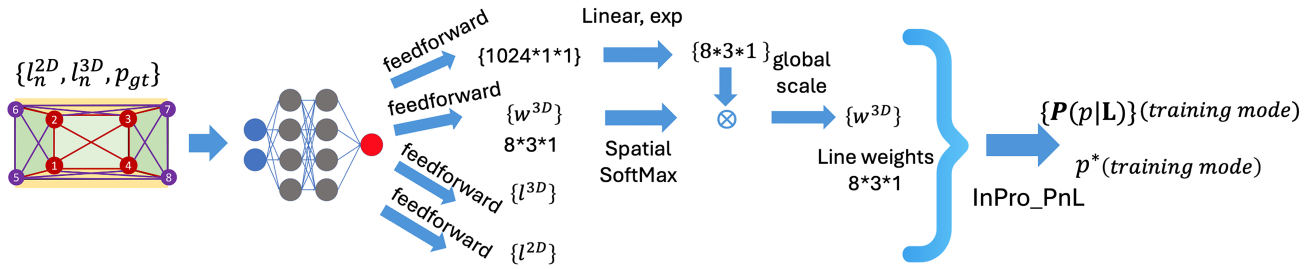


Figure 3. The camera pose estimation network based on MLP with spatial Softmax and global weight scaling.

where $l(.,.)$ represents a pose distance metric. Smooth $L1$ [24] is adopted for position due to its robustness in handling outliers, and cosine similarity is for orientation. Although effective, this regularization loss alone is insufficient for addressing pose ambiguity comprehensively. Therefore, we incorporate it as a supplementary regularization term to enhance training stability and promote convergence [14].

4. Attention-Inspired Correspondence Network

As discussed in Section 3.3, the balance between uncertainty and discriminative power allows identifying important correspondences in a manner analogous to attention mechanisms. This insight motivates us to integrate components from attention-related methods, specifically the Softmax layer and deformable sampling [61]. In this section, we introduce a simple multilayer perceptron (MLP) network [3] incorporating the proposed InPro-PnL layer for indoor camera pose estimation. The primary objective of this network is to demonstrate that the InPro-PnL layer can efficiently and accurately estimate the camera pose by effectively denoising the 2D/3D line correspondences.

In this initial stage of our study, we adopt a standard multilayer perceptron (MLP) network [3] as our baseline architecture, introducing minor modifications to integrate the proposed InPro-PnL layer. Specifically, noisy 2D IICs/IOCs and their corresponding noisy 3D IICs are input to the MLP-based camera pose estimation network, which predicts residual corrections for these correspondences and assigns corresponding weights to each 2D/3D layout line. The primary modifications occur in the output layers, as illustrated in Figure 3. To facilitate stable and efficient training, we incorporate a Softmax operation inspired by attention mechanisms [14, 61], translating absolute correspondence weights into relative importance measures. Additionally, global scaling factors are introduced to regulate the overall concentration of the predicted pose distribution, thereby enhancing the convergence properties of the KL divergence loss. The network is trained using a combination of the KL divergence loss L_{KL} and the regularization loss L_{reg} in Eq. (9), achieving robust performance.

5. Experiments

5.1. Datasets and Metrics

In our experiments, the camera intrinsic matrix is assumed to be known. For evaluating the results across various room layout configurations, we employ the error metrics defined in [15, 16]. Specifically, let $\mathbf{R}_{true}$ and $\mathbf{t}_{true}$ represent the estimated rotation matrix and translation vector, respectively, and let $\mathbf{R}_{true}$ and $\mathbf{t}_{true}$ denote the corresponding ground truth values. Then the rotation error ($Err_{\mathbf{R}}$) and translation error ($Err_{\mathbf{t}}$) are computed as follows:

$$Err_{\mathbf{R}}(deg) = \max_{k\in\{1,2,3\}} \angle(\mathbf{R}_{true}(:,k), \mathbf{R}(:,k)) \times \frac{180}{\pi},$$
$$Err_{\mathbf{t}}(\%) = \frac{||\mathbf{t} - \mathbf{t}_{true}||}{||\mathbf{t}_{true}||} \times 100, \tag{10}$$

For performance evaluation, we compute the mean and median values of the rotation and translation errors. In this study, extensive synthetic datasets for room layout type 0 are generated for training and testing, following the approach described in [16]. Specifically, using a virtual perspective camera with an image resolution of 640×640 pixels and a focal length of 180 pixels, 3D reference lines are created corresponding to various room layout configurations. For each room layout type, the coordinates of the 3D inner corners are fixed within the world coordinate system, along with an initial rotation and translation. We then randomly perturb the rotation angles within the range $[-5, 5]$ and the translation vectors within the range $[-3, 3]$ along three coordinate axes, ensuring that the resultant configurations maintain valid room layouts. Subsequently, these perturbed 3D lines are projected onto the 2D image plane using the ground-truth rotation matrix $\mathbf{R}_{true}$ and translation vector t_{true}. Several randomly generated room layouts are illustrated in Figure 4.

We also evaluate our method using real-world layout images from the Matterport3D dataset [11]; representative examples are shown in Figure 5. Due to time constraints, the

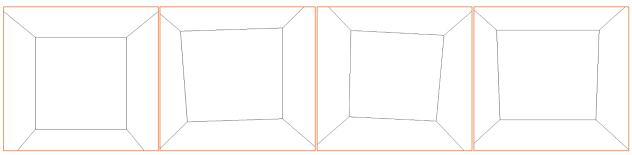


Figure 4. Some randomly generated type 0 room layout images.

current model was trained exclusively on room layout type 0, with other types planned for future studies. All experiments were performed on a workstation running Ubuntu, equipped with an Intel i9 CPU, 64 GB RAM, and a GPU with 24 GB of VRAM.

5.2. Implementation Details

5.2.1. InPro-PnL Configuration

In practice, the orthogonal error terms in Eq. (1) are robustified using the Huber loss function $\xi(\cdot)$ to enhance stability against outliers:

$$\arg\min_{p} \frac{1}{2} \sum_{i=1}^{N} \xi(||f_i(p)||)^2. \tag{11}$$

The Huber loss function with threshold δ is defined as follows:

$$\xi(s) = \begin{cases} s, & s \leq \delta^2, \\ \delta(2\sqrt{s} - \delta), & s > \delta^2. \end{cases} \tag{12}$$

To robustly handle weighted orthogonal errors at varying scales, we employ an adaptive Huber kernel with an object-specific dynamic threshold δ, defined as a function of the correspondence weights $\mathbf{w}_i$ and and the normals of the respective projection plane n_i:

$$\delta = \delta_{\text{rel}} \frac{\|\bar{\mathbf{w}}\|}{2} \left(\frac{1}{N-1} \sum_{i=1}^{N} \|\mathbf{n}_i - \bar{\mathbf{n}}\|^2 \right)^{\frac{1}{2}}, \tag{13}$$

with the relative threshold δ_{rel} as hyperparameter, and the mean vectors

$$\bar{\mathbf{w}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{w}_i, \quad \bar{\mathbf{n}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{n}_i.$$

For the Monte Carlo pose loss, we set the number of AMIS iterations T to 4, with 128 samples per iteration (K' = 128). The loss weights are carefully adjusted so that gradients from L_{KL} match the magnitude of those typically obtained from standard coordinate regression tasks, while gradients from the regularization loss L_{reg} remain significantly smaller. To achieve this balance, we apply the weight normalization technique introduced in [14].

5.2.2. Training the MLP mixed InPro-PnL network

For MLP, we define a single hidden layer with 1024, and set the batch size to 8 for fully training. The network is trained for 10 epochs.

5.3. Results on the synthetic dataset

We conducted training and testing for room layout type 0 using the previously introduced synthetic dataset. Some examples of randomly generated type 0 room layouts are presented in Figure 4. This dataset provides a sufficient sample size, comprising 9,000 data points for training and 1,000 for testing. To evaluate the effectiveness of our proposed network, we compared its performance against state-of-the-art mathematical approaches. Specifically, we benchmarked our method against the PnL-IOC algorithm[16], which has been shown to achieve enhanced accuracy and efficiency, and DLSLU*[52], a method that has demonstrated superior accuracy in type 0 scenarios compared to PnL-IOC. The results demonstrate that the proposed InPro-PnL method exhibits strong noise robustness, as detailed in Table1.

Method	Mean of Rotation error (deg)					
	$\delta = 5$	$\delta = 10$	$\delta = 15$	$\delta = 20$	$\delta = 25$	$\delta = 30$
InPro-PnL	0.5706	0.6620	**0.8211**	**1.0049**	**1.1987**	**1.4188**
PnL-IOC	1.3750	1.8052	2.4779	6.9451	11.0860	20.6360
DLSLU*	**0.3282**	**0.6507**	0.9848	1.4886	1.6164	1.9592
Method	Mean of Translation error (%)					
	$\delta = 5$	$\delta = 10$	$\delta = 15$	$\delta = 20$	$\delta = 25$	$\delta = 30$
InPro-PnL	0.6239	**0.8910**	**1.2082**	**1.5347**	**1.8848**	**2.2803**
PnL-IOC	1.6658	1.9582	5.5359	11.7667	16.3082	26.0812
DLSLU*	**0.6244**	1.1907	1.7448	2.4571	2.8899	3.5188

Table 1. The Rotation error and translation error for type0 with InPro-PnL, PnL-IOC and DLSLU. Experimental results on the simulated data under different noise levels ($\delta = 5, 10, 15, 20, 25, 30$)

The proposed InPro-PnL method exhibits exceptional performance in both rotation and translation error metrics, demonstrating its robustness and reliability in high-noise environments. These results suggest that InPro-PnL effectively integrates room layout information within its geometric framework, enabling more accurate pose estimations compared to existing methods. Furthermore, the consistently low error rates across various noise levels highlight the algorithm's effectiveness and its potential applicability in real-world scenarios where noise and disturbances are prevalent.

5.4. Results on the real image dataset

We initially planned to use the RoomNet network [36] as a baseline and integrate the InPro-PnL layer, aiming to enhance RoomNet's ability to estimate key points for room layout generation while simultaneously obtaining camera pose information based on these key points, which provide room layout line constraints. However, due to time



Figure 5. Some type 0 room layout images.

constraints, we attempted to extract estimated 2D IOCs and IICs for type 0 images (examples shown in Figure 5) from the Matterport3D dataset using the publicly available RoomNet implementation (https://github.com/GitBoSun/roomnet). Unfortunately, the outputs from this network were unreliable and did not yield meaningful results.

As an alternative, we manually extracted 2D IICs and IOCs using MATLAB for 57 type 0 images from the Matterport3D dataset. Due to the occlusion of many IICs and IOCs by furniture, the manual annotation process inevitably introduced significant errors. However, these challenges closely resemble real-world conditions, making the dataset well-suited for evaluating our network. Table 2 provides an example of the extracted data, where some annotated points exhibit minimal error, while those occluded by obstacles tend to have larger deviations. The full evaluation results are presented in Table 3. Despite these challenges, InPro-PnL consistently achieved higher accuracy than other methods when applied to real-world image data.

Ground Truth	Points Marked Manually
(270.2147, 831.7720)	(273.0000, 837.0000)
(1280.0000, 88.5026)	(268.0000, 290.0000)
(1188.4436, 188.0229)	(1196.0000, 228.0000)
(1186.5059, 892.7993)	(1185.0000, 871.0000)
(0.0000, 974.1596)	(1.0000, 963.0000)
(0.0000, 146.9350)	(1.0000, 188.0000)
(1188.4436, 188.0229)	(1279.0000, 149.0000)
(1280.0000, 1022.6701)	(1279.0000, 1009.0000)

Table 2. The ground truth data and error data for 2D IOCs and IICs points .

Method	Mean of Rotation error (deg)	Mean of Translation error (%)
InPro-PnL	**2.7806**	**4.0843**
PnL-IOC	8.3909	10.8736
DLSLU*	5.5870	10.2551

Table 3. The Rotation error and translation error for type0 with InPro-PnL, PnL-IOC and DLSLU*. Experimental results on Matterport3D Dataset with some processed error.

5.5. The Importance of the Softmax Layer

Experimental results show that learning the corresponding weights without the normalization denominator of the spatial Softmax fails to converge.

6. Conclusion and future Work

This study presents InPro-PnL, a novel method that transforms the non-differentiable deterministic PnL operation into a differentiable probabilistic layer, enhancing the flexibility of end-to-end learning for 2D-3D correspondences. This method can be seamlessly integrated into existing deep learning architectures to improve pose estimation accuracy. Specifically, we incorporate the PnL-IOC algorithm within an MLP network for indoor camera pose estimation, demonstrating its effectiveness in mitigating input noise and improving pose accuracy.

Due to time constraints, this study focuses exclusively on room layout type 0. In future work, we aim to extend our deep learning-based approach to additional layout types (2–7) and single-line layouts [36]. The overall framework will remain consistent with that of type 0, with the primary modification involving the adaptation of the PnL solver to accommodate different layout structures. Given the promising results obtained for type 0, we are confident in the method's ability to deliver accurate and reliable results across a broader range of room configurations. Additionally, we plan to evaluate the impact of all hyperparameters and different loss functions related to our networks and select the optimal settings based on experimental results. Furthermore, we intend to integrate the enhanced InPro-PnL framework with RoomNet to further improve room layout estimation accuracy. This integration aims to leverage the complementary strengths of both networks, enabling a more precise interpretation of complex spatial data and enhancing adaptability to dynamic indoor environments.

Moving forward, we also plan to extend this method for camera pose estimation across various room layouts, further refining the accuracy of room layout estimation. Additionally, we will focus on optimizing the entire network by enhancing computational efficiency and refining algorithmic parameters to improve overall performance.

This optimization will involve leveraging advanced machine learning techniques to minimize errors and enhance robustness against diverse environmental conditions.

References

[1] Yousset I Abdel-Aziz, Hauck Michael Karara, and Michael Hauck. Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. *Photogrammetric engineering & remote sensing*, 81(2):103–107, 2015. 2

[2] Adnan Ansar and Konstantinos Daniilidis. Linear pose estimation from points or lines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):578–589, 2003. 2

[3] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003. 5

[4] Christopher M Bishop. Mixture density networks. 1994. 3

[5] Eric Brachmann and Carsten Rother. Learning less is more-6d camera localization via 3d surface regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4654–4662, 2018. 2

[6] Eric Brachmann and Carsten Rother. Visual camera relocalization from rgb and rgb-d images using dsac. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5847–5865, 2021. 3

[7] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, et al. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3364–3372, 2016. 3

[8] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6684–6692, 2017. 2

[9] Mai Bui, Tolga Birdal, Haowen Deng, Shadi Albarqouni, Leonidas Guibas, Slobodan Ilic, and Nassir Navab. 6d camera relocalization in ambiguous scenes via continuous multimodal inference. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 139–157. Springer, 2020. 3

[10] Dylan Campbell, Liu Liu, and Stephen Gould. Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 244–261. Springer, 2020. 2, 3

[11] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 5

[12] Bo Chen, Alvaro Parra, Jiewei Cao, Nan Li, and Tat-Jun Chin. End-to-end learnable geometric vision by backpropagating pnp optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8100–8109, 2020. 2, 3

[13] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10379–10388, 2021. 2

[14] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2781–2790, 2022. 1, 3, 4, 5, 6

[15] Xiaowei Chen and Guoliang Fan. Egocentric indoor localization from coplanar two-line room layouts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1549–1559, 2022. 1, 5

[16] Xiaowei Chen and Guoliang Fan. Indoor camera pose estimation from room layouts and image outer corners. *IEEE transactions on multimedia*, 25:7992–8005, 2023. 1, 2, 4, 5, 6

[17] Stéphane Christy and Radu Horaud. Iterative pose computation from line correspondences. *Computer vision and image understanding*, 73(1):137–144, 1999. 2

[18] JEAN-MARIE CORNUET, JEAN-MICHEL MARIN, Antonietta Mira, and Christian P Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39 (4):798–812, 2012. 4

[19] James M Coughlan and Alan L Yuille. Manhattan world: Orientation and outlier detection by bayesian inference. *Neural computation*, 15(5):1063–1088, 2003. 1

[20] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2

[21] Haowen Deng, Mai Bui, Nassir Navab, Leonidas Guibas, Slobodan Ilic, and Tolga Birdal. Deep bingham networks: Dealing with uncertainty and ambiguity in pose estimation. *International Journal of Computer Vision*, 130(7): 1627–1654, 2022. 3

[22] Fadi Dornaika and Christophe Garcia. Pose estimation using point and line correspondences. *Real-Time Imaging*, 5(3): 215–230, 1999. 2

[23] Igor Gilitschenski, Roshni Sahoo, Wilko Schwarting, Alexander Amini, Sertac Karaman, and Daniela Rus. Deep orientation uncertainty learning based on a bingham loss. In *International conference on learning representations*, 2019. 2

[24] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 5

[25] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2

[26] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2888–2897, 2019. 2

[27] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. In *2009 IEEE 12th international conference on computer vision*, pages 1849–1856. IEEE, 2009. 2

[28] Yaxuan Huang, Xili Dai, Jianan Wang, Xianbiao Qi, Yixing Yuan, and Xiangyu Yue. Unposed sparse views room layout reconstruction in the age of pretrain model. *arXiv preprint arXiv:2502.16779*, 2025. 2

[29] Shun Iwase, Xingyu Liu, Rawal Khirodkar, Rio Yokota, and Kris M Kitani. Repose: Fast 6d object pose refinement via deep texture rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3303–3312, 2021. 2

[30] Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, and Ming Zhao. Lgt-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1654–1663, 2022. 2

[31] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 2

[32] Minjung Kim, Junseo Koo, and Gunhee Kim. Ep2p-loc: End-to-end 3d point to 2d pixel localization for large-scale visual localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21527–21537, 2023. 1

[33] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 2

[34] Rakesh Kumar and Allen R Hanson. Robust methods for estimating pose and a sensitivity analysis. *CVGIP: Image understanding*, 60(3):313–342, 1994. 2

[35] Mohammed Kutbi, Haoxiang Li, Yizhe Chang, Bo Sun, Xin Li, Changjiang Cai, Nikolaos Agadakos, Gang Hua, and Philippos Mordohai. Egocentric computer vision for hands-free robotic wheelchair navigation. *Journal of Intelligent & Robotic Systems*, 107(1):10, 2023. 1

[36] Chen-Yu Lee, Vijay Badrinarayanan, Tomasz Malisiewicz, and Andrew Rabinovich. Roomnet: End-to-end room layout estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 4865–4874, 2017. 1, 6, 7

[37] Hung Jin Lin and Shang-Hong Lai. Deeproom: 3d room layout and pose estimation from a single image. In *Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26–29, 2019, Revised Selected Papers, Part II 5*, pages 719–733. Springer, 2020. 1, 2

[38] Osama Makansi, Eddy Ilg, Ozgun Cicek, and Thomas 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*, pages 7144–7153, 2019. 3

[39] Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Relative camera pose estimation using convolutional neural networks. In *Advanced Concepts for Intelligent Vision Systems: 18th International Conference, ACIVS 2017, Antwerp, Belgium, September 18-21, 2017, Proceedings 18*, pages 675–687. Springer, 2017. 3

[40] Faraz M Mirzaei and Stergios I Roumeliotis. Globally optimal pose estimation from line correspondences. In *2011 IEEE International Conference on Robotics and Automation*, pages 5581–5588. IEEE, 2011. 2

[41] Narendra Mohan and Manoj Kumar. Room layout estimation in indoor environment: a review. *Multimedia Tools and Applications*, 81(2):1921–1951, 2022. 1, 2

[42] Luca Del Pero, Joshua Bowdish, Daniel Fried, Bonnie Kermgard, Emily Hartley, and Kobus Barnard. Bayesian geometric modeling of indoor scenes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2719–2726. IEEE, 2012. 2

[43] Luca Del Pero, Joshua Bowdish, Bonnie Kermgard, Emily Hartley, and Kobus Barnard. Understanding bayesian rooms using composite 3d object models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 153–160, 2013. 2

[44] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *International Journal of Computer Vision*, 132(11):4880–4936, 2024. 1

[45] Bronislav Přibyl, Pavel Zemčík, and Martin Čadík. Camera pose estimation from lines using pl\" ucker coordinates. *arXiv preprint arXiv:1608.02824*, 2016. 2

[46] Jianing Qiu, Frank P-W Lo, Xiao Gu, Yingnan Sun, Shuo Jiang, and Benny Lo. Indoor future person localization from an egocentric wearable camera. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8586–8592. IEEE, 2021. 1

[47] Francesco Ragusa, Antonino Furnari, Sebastiano Battiato, Giovanni Signorello, and Giovanni Maria Farinella. Egocentric visitors localization in cultural sites. *Journal on Computing and Cultural Heritage (JOCCH)*, 12(2):1–19, 2019. 1

[48] Alexander G Schwing, Tamir Hazan, Marc Pollefeys, and Raquel Urtasun. Efficient structured prediction for 3d indoor scene understanding. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2815–2822. IEEE, 2012. 2

[49] Yoli Shavit and Ron Ferens. Introduction to camera pose estimation with deep learning. *arXiv preprint arXiv:1907.05272*, 2019. 3

[50] Manuel Silva, Ricardo Ferreira, and José Gaspar. Camera calibration using a color-depth camera: Points and lines based dlt including radial distortion. In *Workshop Color-Depth Camera Fusion in Robotics, held with IROS*, 2012. 2

[51] David E Tyler. Statistical analysis for the angular central gaussian distribution on the sphere. *Biometrika*, 74(3):579–589, 1987. 4

[52] Alexander Vakhitov, Luis Ferraz, Antonio Agudo, and Francesc Moreno-Noguer. Uncertainty-aware camera pose estimation from points and lines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4659–4668, 2021. 2, 6

[53] Huayan Wang, Stephen Gould, and Daphne Roller. Discriminative learning with latent variables for cluttered indoor scene understanding. *Communications of the ACM*, 56(4): 92–99, 2013. 2

[54] Ping Wang, Guili Xu, Yuehua Cheng, and Qida Yu. Camera pose estimation from lines: a fast, robust and general method. *Machine Vision and Applications*, 30:603–614, 2019. 2

[55] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1–10, 2020. 2

[56] Chi Xu, Lilian Zhang, Li Cheng, and Reinhard Koch. Pose estimation from line correspondences: A complete analysis and a series of solutions. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1209–1222, 2016. 1, 2

[57] Chenggang Yan, Biyao Shao, Hao Zhao, Ruixin Ning, Yongdong Zhang, and Feng Xu. 3d room layout estimation from a single rgb image. *IEEE Transactions on Multimedia*, 22 (11):3014–3024, 2020. 1, 2

[58] Lilian Zhang, Chi Xu, Kok-Meng Lee, and Reinhard Koch. Robust and efficient pose estimation from line correspondences. In *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part III 11*, pages 217–230. Springer, 2013. 2

[59] Weidong Zhang, Wei Zhang, Kan Liu, and Jason Gu. Learning to predict high-quality edge maps for room layout estimation. *IEEE Transactions on Multimedia*, 19(5):935–943, 2016. 2

[60] Weidong Zhang, Wei Zhang, and Yinda Zhang. Geolayout: Geometry driven room layout estimation based on depth maps of planes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 632–648. Springer, 2020. 2

[61] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5

[62] Shengbin Zhuang, Zongmin Zhao, Lin Cao, Dongfeng Wang, Chong Fu, and Kangning Du. A robust and fast method to the perspective-n-point problem for camera pose estimation. *IEEE Sensors Journal*, 23(11):11892–11906, 2023. 1