Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Approach to Robust Visual Relocalization for Humanoid Robots via Mixture-of-Experts with Hierarchical Distillation

Shichu Sun
Yunnan Normal University
Jingwen Luo
by1503117@buaa.edu.cn

Yunnan Normal University

Research Article

Keywords: Visual relocalization, Humanoid robots, Mixture-of-Experts (MoE), Hierarchical distillation

Posted Date: September 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8936961/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Approach to Robust Visual Relocalization for Humanoid Robots via Mixture-of-Experts with Hierarchical Distillation

Shichu Sun[1] and Jingwen Luo[1*]

[1]School of Information Science and Technology, Yunnan Normal University, No. 768 Juxian Street, Chenggong District, Kunming, 650500, China.

*Corresponding author(s). E-mail(s): by1503117@buaa.edu.cn;
Contributing authors: 2324100044@ynnu.edu.cn;

Abstract

Humanoid robots are prone to interference from factors such as abrupt changes in angle-of-view and self-occlusion during operation, making it difficult for their visual relocalization to balance accuracy, stability, and real-time performance. To this end, we propose a robust visual relocalization framework integrating Mixture-of-Experts (MoE) with hierarchical knowledge distillation. The framework utilizes a point–line scene coordinate mapping network (PL2Map) regressor to map the extracted point-line features from scene images to a unified representation space, so as to obtain robust feature representations under multi scene conditions. Also, a Soft-MoE with sparse gating is constructed in the intermediate layer to provide reasonable inference paths for inputs of different scene structures. On this basis, an expert selection-aware hierarchical distillation(ESHD) is developed. During training, multi-level distillation over feature representations, routing distributions, and geometric outputs aligns the teacher model "MoE-PL2Map" with the student model, achieving substantial parameter compression while maintaining performance comparable to the teacher; at inference, the 6 degrees of freedom (6-DoF) camera pose is recovered via scene coordinate based pose regression and geometric consistency constraints. Extensive simulation studies on public datasets and real-world experiments on the "KUAVO-4pro" humanoid robot demonstrate that the proposed method can achieve accurate and efficient robust relocalization under various scene conditions.

Keywords: Visual relocalization, Humanoid robots, Mixture-of-Experts (MoE), Hierarchical distillation

1 Introduction

Visual relocalization plays a pivotal role in computer vision and embodied robotics. It is fundamental for humanoid robots, autonomous driving systems [1], and augmented reality devices [2] to achieve stable environmental perception, precise decision-making, and autonomous behavior in complex dynamic real-world environments. For humanoid robots with high degrees of freedom, the 6-DoF camera pose must be estimated in real time under challenging conditions such as abrupt illumination changes, dynamic occlusions, texture-poor regions, and structural changes [3], all under limited computational resources. However, during long-term operation, pose drift induced by environmental disturbances and robot self-motion accumulates over time. Consequently, achieving high-accuracy while simultaneously maintaining robustness and real-time performance remains a core challenge for visual relocalization on humanoid platforms [4].

Studies [5] has shown that existing visual relocalization schemes struggle to produce accurate pose estimates in scenarios with large angle-of-view changes and generally lack self-assessment mechanisms for relocalization accuracy, which limits both recall and precision. This reveals a persistent difficulty in jointly balancing feature quality, cross-scene generalization, and computational cost. Specifically, geometry-based methods that rely on handcrafted features and prioritized matching perform well in texture-rich environments, but their performance degrades drastically in low-texture or dynamic scenes, or when matching quality deteriorates [6]. Purely learned regression models, such as PoseNet [7], possessed strong representation capacity but lack explicit geometric constraints, making it difficult for their accuracy to surpass retrieval-based or 3D structure–based approaches [8, 9]. Brachmann et al. [10, 11] improved robustness by incorporating geometric consistency into learned models, yet these methods typically incur high model complexity and inference cost, making them unsuitable for the real-time constraints of robotic platforms [5, 9]. Recently, image-retrieval-based methods such as Random Ferns [12] have been widely adopted in SLAM systems due to their low memory footprint and real-time matching capability. However, they struggle to generalize to

novel viewpoints and cannot self-assess pose accuracy [5, 12].

To address these limitations, this paper proposes a visual relocalization framework that combines a MoE-based architecture with an expert selection-aware hierarchical distillation(ESHD) strategy. We introduce sparsely activated expert routing [13] into the feature regression module, enabling input-dependent allocation of complementary expert subnetworks for different geometric and semantic patterns and thereby enhancing structural adaptability under a fixed computational budget. Meanwhile, expert-selection-aware distillation incorporates both intermediate feature representations and expert routing behavior into teacher–student alignment, enabling a lightweight student model to preserve the teacher network's semantic expressiveness and structural selection behavior while substantially reducing the parameter count [14, 15]. At the level of the scene's geometric structure, we construct a geometry-consistent pose estimation module by combining point–line reprojection constraints, which improves stability and interpretability in weakly textured and structurally degraded scenes [16, 17]. The main contributions of this work are summarized as follows:

(1) A unified point–line scene coordinate mapping framework, PL2Map, is proposed for visual relocalization of humanoid robots, enhancing relocalization accuracy and robustness across diverse environmental conditions.
(2) A Soft-MoE with sparse gating is constructed within the intermediate geometric feature space of PL2Map. It differentiates features of different scenes through conditional routing, thereby improving the model's adaptability under various environmental conditions with limited computational resources.
(3) A MoE-ESHD mechanism is developed to align the teacher model MoE-PL2Map with the student model at multiple levels of feature representation, routing distribution, and geometric output, allowing the student model to maintain consistent structural behavior and semantic representation with the teacher model while significantly compressing parameters.

The remainder of this paper is organized as follows. Section 2 briefly reviews related work on visual relocalization. Section 3 provides a general description of the characteristics of humanoid robots. Section 4 presents an overview of the proposed framework and details of the methodology. Section 5 reports representative simulation studies and experimental results on a humanoid robot platform. Finally, Section 6 concludes the paper and outlines directions for future work.

2 Related Work

2.1 Retrieval-based methods

Typically, image retrieval methods rely on local features and geometric constraints to realized relocalization. Sattler et al. [6] achieved high-accuracy in texture-rich scenes using a structured pipeline based on SIFT/SURF/ORB and PnP-RANSAC. Glocker et al. [12] achieved RGB-D camera relocalization by encoding keyframes using random ferns, demonstrating outstanding performance in both memory efficiency and real-time capability. NetVLAD, proposed by Arandjelović et al. [18], effectively enhanced the expressive power for images within the same category, and its global descriptor effectively reduced matching complexity in large-scale scenes through image retrieval. However, such methods are prone to feature degradation in areas with strong lighting changes, dynamic occlusion, and weak texture, resulting in sparse or even ineffective features in 2D-3D matching. Moreover, retrieval-based schemes inherently rely on appearance similarity, limiting their adaptability to long-term appearance drift and large angle-of-view changes. In summary, retrieval-based methods remain advantageous in environments with well-defined structures and stable textures, but for humanoid robots performing long-term autonomous operation in complex dynamic scenes, their robustness and generalization capabilities are still insufficient.

2.2 Scene-coordinate-regression-based methods

Regression-based visual relocalization aims to directly learn the mapping from images to poses using deep networks. Absolute Pose Regression (APR) methods, exemplified by PoseNet [7], employed convolutional networks to regress 6-DoF poses in an end-to-end manner, significantly simplifying feature extraction and geometric solving. However, subsequent studies have shown that the lack of explicit geometric constraints often limits their accuracy compared with structured approaches [8, 9]. Scene Coordinate Regression (SCoRe) enhanced geometric consistency by predicting a 3D scene coordinate for each pixel and supervising the model using a PnP-based closed loop. The regression forest proposed by Shotton et al. [19] and the DSAC series introduced by Brachmann et al. [10, 11] further reformulate RANSAC as a differentiable process, enabling joint optimization of scene coordinate prediction and interior point selection. These approaches achieved accuracy comparable to or even surpassing structured methods on standard indoor datasets. Nevertheless, APR and SCoRe models typically contain a large number of parameters, incur high training costs, and remain sensitive to noise and domain shifts. As a result, when deployed on resource-constrained humanoid robot, it is still difficult to balance representational capacity and computational overhead [20].

2.3 Multimodal-enhancement-based methods

To enhance the stability in complex scenes and long sequences, some researchers have incorporated temporal modeling and multimodal information augmentation into single-frame relocalization frameworks. Walch et al. [21] drew on LSTMs to perform temporal modeling on APR outputs. Valentin et al. [22] and Lu et al. [5] explicitly modeled prediction uncertainty and self-assessed accuracy using regression forests and hierarchical Fern encodings, respectively, and mitigated drift and mismatches through closed-loop correction. Cavallari et al. [23] proposed a cascaded relocalization scheme that balances accuracy and real-time performance via a coarse-to-fine multi-stage architecture. In addition, methods such as InLoc [24] strengthened visibility reasoning through dense matching and virtual view synthesis. Zhou et al. [25] leveraged semantic cues to alleviate ambiguities caused by weak textures and repetitive structures, while self-supervised features such as SuperPoint [26] exhibited better stability across perspectives and lighting conditions. Overall, these approaches illustrate a clear trend from single-frame, single-modality methods toward temporal constraints and multi-source information fusion. Unfortunately, additional sequence modeling and modality branches inevitably increase model complexity and inference latency, which remains problematic for embedded humanoid robot platforms with strict real-time and resource constraints.

2.4 MoE-based methods

For the diversity of scenes and complex input patterns faced by visual relocalization, a single-path network is often insufficient to balance different geometric patterns and appearance distributions. The Mixture of Experts (MoE) model provides a conditional-computation-based structural adaptation mechanism for this purpose. The sparse-gated MoE architectures proposed by Shazeer et al. [13] and Fedus et al. [27] use a gating network to perform sparse routing among multiple experts, allowing the model to expand its effective capacity under an approximately fixed computational budget. Riquelme et al. [28] further introduced MoE into vision Transformers (ViT), spatially activating different experts across image regions to enhance representational diversity. Subsequently, the application of MoE has also expanded to robustness-oriented and resource-constrained scenarios. Pavlitska et al. [29] systematically analyzed the robustness of MoE from the perspective of adversarial training and proposed an adaptive adversarial training and dual-model collaboration strategy, which significantly improves robustness to adversarial perturbations while maintaining accuracy. In addition, for resource-constrained deployment, Gao et al. [30] proposed the AnyExperts framework, a budget-aware dynamic routing strategy for sparse MoE that adaptively selects experts under explicit compute budgets, providing a new design paradigm for improving MoE inference efficiency under strict computational constraints. Overall, MoE offers an effective means of handling diverse input distributions and increasing model capacity under limited computational resources; however, most existing studies focus on classification and general vision or multimodal tasks, and systematic investigations of MoE in geometry-sensitive visual relocalization remain scarce.

2.5 Knowledge-distillation-based methods

Complementary to the structural adaptability of MoE, knowledge distillation (KD) aims to transfer the representational capacity of high-capacity teacher networks to compact student models while keeping inference cost under control. Hinton et al. [14] introduced distillation based on output distributions using soft labels, and Romero et al. [15] extended this idea to intermediate feature maps. Park [31] and Tung [32] developed structure-preserving distillation strategies that exploit inter-sample relationships and feature-space geometry, enabling student networks to approximate teachers at multiple representation levels. Moslemi et al. [33] systematically reviewed recent advancements in knowledge distillation and pointed out that KD has become a key approach to improving model performance and generalization, including in resource-constrained deployment scenarios [34]. Qian et al.'s MLEMKD [35] enhanced teacher-student information exchange through mutual learning, extending KD to temporal reasoning tasks. Feng et al. [36] proposed a route-aware distillation framework based on a mixture of LoRA experts, where expert selection information is explicitly incorporated into the distillation targets for large language model based bundle generation. Although these works significantly enrich the forms and applications of KD, most of them target classification or knowledge reasoning tasks. Even in MoE-Distill, routing-aware distillation is mainly used to improve discriminative performance, and there is limited systematic modeling of the coupling between expert selection and geometric representation in geometric regression tasks.

In general, visual relocalization has evolved from structured and retrieval-based methods reliant on geometric feature matching to integrated frameworks combining deep regression, differentiable geometric reasoning, and temporal and multimodal enhancements. It has begun incorporating conditional computation and model compression techniques such as MoE and knowledge distillation to improve accuracy and robustness under limited computational resources. Nonetheless, geometry-based, regression-based, and multimodal-enhancement-based methods still face distinct limitations in complex dynamic scenes, including feature degradation, imbalance between generalization and computational cost, and increased latency and implementation complexity [5]. Motivated by these challenges, the proposed MoE-ESHD framework combines sparse expert selection with ESHD under point–line geometric consistency constraints, enabling visual relocalization on humanoid robot platforms that

jointly satisfies requirements on accuracy, robustness, and real-time performance.

3 Characteristics of humanoid robot

Humanoid robots mimic human bones and joint structures, with multiple movable joints that can achieve flexible motion and posture control in various terrain conditions. This allows their base state to continuously change during motion, without a fixed reference position, making them a floating reference robot [37]. Fig. 1 shows the KUAVO-4pro humanoid robot (1.66m high, 55kg heavy) involved in this work, which has differences in visual input characteristics and motion modes compared to traditional mobile robots [38].

On one hand, as the use of a multi-degree-of-freedom bionic structure (the entire body has 22 degrees of freedom (22-DoF) excluding the end effector, with a single leg alone possessing 6-DoF), the head-mounted camera experiences significant short-term positional and postural changes under the continuous influence of gait cycles, torso posture adjustments, and upper limb operations. Consequently, image sequences frequently exhibit rotational jumps, blurring, and field-of-view(FoV) shifts. During tasks such as grasping, the movements of the robotic arm and hand frequently enter the FoV, causing partial occlusions that result in discontinuous and unstable feature distributions; On the other hand, typical working environments for robots include corridors, equipment rooms, and outdoor areas with complex lighting conditions. These environments often have weak textures, repetitive structures, and mirror and transparent materials, which makes it easy for relocation methods that rely on a single type of feature to perform unevenly in different areas and difficult to maintain stability between texture dominated and structure dominated scenes. Additionally, the onboard processing unit of the robot is constrained by computational power, memory, and power consumption. The localization module must complete feature encoding and pose estimation within a limited time budget, making it impossible to rely on complex optimization processes or large-scale network architectures.

4 Framework and methodology

In this work, addressing the characteristics of humanoid robots and practical application requirements in real-world scenes, we propose a visual relocalization framework that integrates a MoE model with hierarchical distillation, as illustrated in Fig. 2. First, SuperPoint [26] and DeepLSD [39] are employed at the front-end of the constructed point–line scene coordinate mapping network (PL2Map) scene coordinate regression network to extract point and line features respectively. Subsequently, a MoE-Attentional GNN architecture is incorporated within the intermediate geometric feature space of PL2Map to adaptively model diverse geometric patterns, thereby achieving

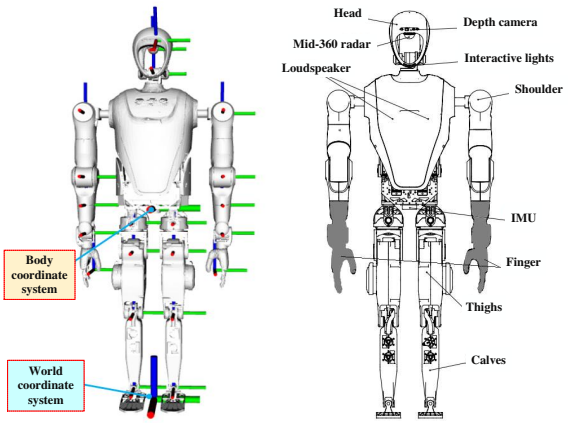


Fig. 1 KUAVO-4pro humanoid robot.

robust relocalisation through point-line 3D regression and geometric reprojection loss. Furthermore, we develop an ESHD strategy to transfer the geometric structural behavior of the MoE-PL2Map teacher network to a lightweight student network, thus achieving efficient deployment on humanoid robots.

4.1 Feature extraction and encoding

To extract local texture and corner clues from scene images, we employ the SuperPoint network [26] to extract a set of keypoints $\{p_i\}_{i=1}^{N_p}$ and their 256-dimensional descriptors $\{d_i^{(p)}\}_{i=1}^{N_p}$. Also, DeepLSD [39] is utilised to detect a set of line segments $\{l_j\}_{j=1}^{N_l}$ and their endpoints $\{a_j, b_j\}$, thereby providing scale structure boundaries. To incorporate line segments into a unified learning framework, we discretize each line segment by sampling n equidistant points between the endpoints a_j and b_j. In this way, each line segment is represented by $n+2$ discrete points (including endpoints a_j and b_j), i.e.

$$p_{j,k} = a_j + \frac{k}{n+1}(b_j - a_j), \quad k = 1, \ldots, n \tag{1}$$

Subsequently, these discrete points along the lines, together with the key points, are mapped onto SuperPoint's feature map to extract the corresponding local descriptors. Let the local descriptor vector for keypoint p_i be denoted as $d_i^{(p)} \in \mathbb{R}^{256}$. For each line segment l_j, the initial descriptor set for its sampled points is denoted as $d_i^{(p)} \in \mathbb{R}^{256}$. By applying a line segment geometric encoder to aggregate and encode this set, the descriptor vector $d_j^{(l)} \in \mathbb{R}^D$ for the line segment is obtained. In our case, the dimensionality of the line segment descriptor is maintained consistent with that of the point descriptor, i.e. $D = 256$ thereby yielding a set of point features and line features that are both heterogeneous and dimensionally consistent.

To further unify the expression of heterogeneous features into the geometric encoding space of PL2Map, point and line descriptors are embedded separately through learnable linear mappings, i.e.

$$e_i^{(p)} = W_p d_i^{(p)}, \qquad e_j^{(l)} = W_l d_j^{(l)} \tag{2}$$

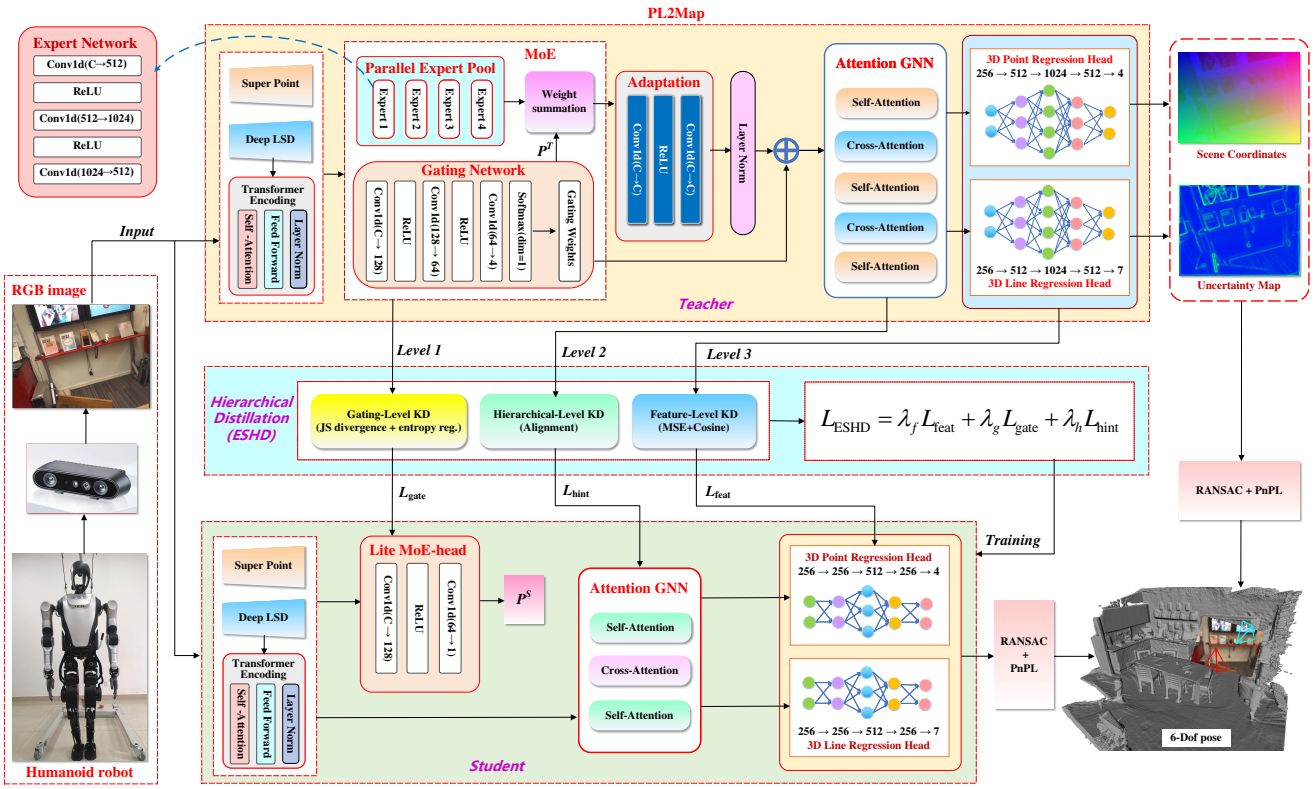


Fig. 2 Scheme of the proposed robust visual relocalisation framework for humanoid robots via MoE with hierarchical distillation. The teacher adopts a PL2Map backbone with a Soft MoE-Attentional GNN head, while the student uses a compact single-path head and ESHD to inherit the teacher's structural behaviours.

where $W_p, W_l \in \mathbb{R}^{C\times D}$ are two learnable matrices that respectively map the original point and line descriptors $d_i^{(p)}, d_j^{(l)} \in \mathbb{R}^D$ onto a unified C-dimensional geometric space to yield $e_i^{(p)}, e_j^{(l)} \in \mathbb{R}^C$. This yields the embedded point feature set $\{e_i^{(p)}\}_{i=1}^{N_p}$ and the line feature set $\{e_j^{(l)}\}_{j=1}^{N_l}$, which serve as the initial representation for the intermediate geometric encoding in PL2Map.

As a notes, the point descriptor output by SuperPoint can be directly regarded as the embedded point feature $e_i^{(p)}$, and the line segment descriptor is encoded by LineEncoder to obtain $e_i^{(l)}$ through Transformer encoding. Consequently, the mapping W_p in the point branch is approximately an identity transformation, while W_l in the line branch is implicitly realized by the line-segment geometric encoder.

4.2 MoE-Attentional GNN

In practical scenes, the spatial distribution of local textures, structural contours, and dynamic occlusions exhibits significant heterogeneity, making it challenging for a single model to simultaneously capture the geometric structures of diverse scenes. To adaptively model distinct geometric patterns in the point-line geometric feature space, we introduce a MoE framework in the intermediate feature space between the point branch and line branch of PL2Map, and construct a set of gate networks and expert subnetworks for each of them. Through a conditional computation mechanism to dynamically select the most suitable expert subnetwork for the current input during the inference stage, thus achieving adaptive modeling for scene geometry characteristics. For the point branch, let $e_i^{(p)}$ denote the input feature of the i^{th} point, a gating network $g_p(\cdot)$ computes the logit $z_{p,i}^{(k)}$ associated with the k^{th} expert for this point as follows:

$$z_{p,i}^{(k)} = g_p^{(k)}\big(e_i^{(p)}\big), \qquad k = 1, \ldots, E \tag{3}$$

where E is the number of experts (in this work $E = 4$). Further, the Softmax function is used to convert it into the selection probabilities of each expert, i.e.

$$\alpha_{p,i}^{(k)} = \frac{\exp\big(z_{p,i}^{(k)}\big)}{\sum_{s=1}^{E} \exp\big(z_{p,i}^{(s)}\big)} \tag{4}$$

where E is the total number of experts, and $E = 4$ in this paper.Let the k^{th} expert subnetwork of the point-based branch be denoted as $E_(p)^k(\cdot)$. Then, the output of the point after undergoing the MoE transformation is as follows:

$$\tilde{e}_i^{(p)} = \sum_{k=1}^{E} \alpha_{p,i}^{(k)}\, E_k^{(p)}\big(e_i^{(p)}\big) \tag{5}$$

To maintain consistency with the original feature space, the features output by the MoE undergo a lightweight feature adaptation network $a_p(\cdot)$ and are connected to the input features via a residual connection, thereby yielding the updated point features as follows:

$$h_i^{(p,0)} = e_i^{(p)} + a_p\big(\tilde{e}_i^{(p)}\big) \tag{6}$$

Additionally, stacking the gating probabilities of all points by column yields the following gating weight matrix for the point branch, i.e.

$$G_p = \left[\, \alpha_{p,1}, \ldots, \alpha_{p,N_p} \,\right] \in \mathbb{R}^{E \times N_p} \tag{7}$$

Likewise, in terms of the line branch input feature $e_j^{(l)}$, the gating network $g_l(\cdot)$, and the expert $\{E_k^{(l)}(\cdot)\}_{k=1}^{E}$, the updated line feature $h_j^{l,(0)}$ and the gating weight matrix $G_l \in \mathbb{R}^{E \times N_l}$ for the line branch are obtained, serving as input for subsequent gate-level distillation in the ESHD.

To explicitly model the geometric relationships between points and lines, we further apply an Attentional Graph Neural Network (Attentional GNN) on top of the MoE outputs to jointly encode geometric context. The Attentional GNN is constructed as a stack of propagation blocks with a "self–cross–self–cross–self" topology, and the point node update in the l^{th} layer can be written as follows:

$$\begin{aligned} h_i^{p,(l+1)} = h_i^{p,(l)} &+ \Delta_{\text{self}}^{p}\big(h_i^{p,(l)}, \mathcal{N}_i^{p}\big) \\ &+ \Delta_{\text{cross}}^{p}\big(h_i^{p,(l)}, \{\, h_j^{l,(l)} \,\}_{j=1}^{N_l}\big) \end{aligned} \tag{8}$$

where $\mathcal{N}_i^p$ denotes the neighbourhood of point i, Δ_{self}^p and Δ_{cross}^p respectively model point-point and point-line geometric relationships through self-attention and point-line cross-attention. Similarly, the line node at l^{th} layer is updated as follows:

$$\begin{aligned} h_j^{l,(l+1)} = h_j^{l,(l)} &+ \Delta_{\text{self}}^{l}\big(h_j^{l,(l)}, \mathcal{N}_j^{l}\big) \\ &+ \Delta_{\text{cross}}^{l}\big(h_j^{l,(l)}, \{\, h_i^{p,(l)} \,\}_{i=1}^{N_p}\big) \end{aligned} \tag{9}$$

where $\mathcal{N}_j^l$ is the neighbourhood of line segment j.

Through the stacking of a multi-layer "self– cross –self–cross–self" architecture, point features and line features repeatedly interact within multi-scale neighborhoods, progressively encoding local textures, structural contours, and global geometric constraints. After L-layer propagation, the context enhanced point set $\{h_i^{p,(L)}\}_{i=1}^{N_p}$ and line set $\{h_j^{l,(L)}\}_{j=1}^{N_l}$ are obtained.

The resulting Soft MoE-Attentional GNN intermediate module thus enjoys two benefits: on the one hand, the gating network adaptively distributes computation across expert subnetworks, achieving an expressive capacity comparable to an ensemble of experts. On the other hand, explicit modeling of point–line relations enhance geometric structure awareness. Notably, in this work, we did not use additional complex training techniques such as load balancing regularization, and only used uniform ℓ_2 gradient pruning [40] as the basic numerical stabilization method to enable the model to converge stably under standard training configurations.

4.3 Scene coordinate regression

At the output of the PL2Map framework, we use point-line scene coordinate regression to map the structured geometric features $\{h_i^{p,(L)}\}$ and $\{h_j^{l,(L)}\}$ output by the MoE-Attention GNN module to 3D scene coordinates, and apply geometric consistency constraints through explicit reprojection errors. Then, in the inference stage, we utilize a PnP solver with RANSAC to obtain the 6-DoF pose of the camera. For the i^{th} context-enhanced point feature $h_i^{p,(L)}$, the point-branch regression head $f_p(\cdot)$ outputs a 4-dimensional vector, i.e.

$$\big[\hat{P}_i, \hat{\sigma}_i^{(p)}\big]^{\top} = f_p\big(h_i^{p,(L)}\big), \qquad \hat{P}_i \in \mathbb{R}^3 \tag{10}$$

where $\hat{P}_i$ denotes the predicted 3D scene coordinate and $\hat{\sigma}_i^{(p)}$ denotes an uncertainty parameter associated with this point. For the line branch,$f_l(\cdot)$ takes the line feature $h_j^{l,(L)}$ as input and outputs the two endpoints of the line segment in 3D space and their uncertainties, i.e.

$$\big[\hat{A}_j, \hat{B}_j, \hat{\sigma}_j^{(l)}\big]^{\top} = f_l\big(h_j^{l,(L)}\big), \qquad \hat{A}_j, \hat{B}_j \in \mathbb{R}^3 \tag{11}$$

In our case, f_p and f_l are two multi-layer perceptrons (MLPs) with output dimensions 4 and 7, respectively.During training, let the true 3D scene coordinates of each point be P^* , and the true 3D endpoints of each line segment be A_j^* and B_j^*.Then the scene coordinate regression loss is defined as follows:

$$L_{\text{3D}}^{p} = \frac{1}{B} \sum_{b=1}^{B} \sum_{i} m_{b,i}^{p} \big\| \hat{P}_{b,i} - P_{b,i}^{*} \big\|_2 \tag{12}$$

$$L_{\text{3D}}^{l} = \frac{1}{B} \sum_{b=1}^{B} \sum_{j} m_{b,j}^{l} \Big\| \big[\hat{A}_{b,j}, \hat{B}_{b,j}\big] - \big[A_{b,j}^{*}, B_{b,j}^{*}\big] \Big\|_2 \tag{13}$$

where B is the size of batch, $m_{b,i}^{(p)}, m_{b,j}^{(l)} \in \{0,1\}$ are validity masks indicating whether the corresponding point or line has reliable ground truth. In parallel, the network uses the predicted uncertainty parameters $\hat{\sigma}_i^{(p)}$ and $\hat{\sigma}_j^{(l)}$ to derive confidence scores $c_i^{(p)}, c_j^{(l)} \in (0,1)$ through a monotonic mapping function. These confidences are encouraged to approximate the validity masks, yielding uncertainty regularization terms L_{unc}^{p} and L_{unc}^{l} , which encourage the model to automatically reduce confidence in low-light or weak-texture regions and thereby improve the robustness.

To further enhance consistency between the predicted scene coordinates and the observed image measurements, we introduce explicit point-line reprojection errors during training. Namely, given the camera intrinsic matrix K and the ground-truth pose (R, t), We project the 3D points and 3D line segments predicted by the network back onto the image plane to align with the original observed features, and minimize the corresponding reprojection error, thereby explicitly applying geometric consistency constraints during training. For each predicted 3D point P_i, the projection is writtren as follows:

$$\tilde{p}_i = \Pi\big(R\hat{P}_i + t,\, K\big) \tag{14}$$

where $\Pi(\cdot)$ denotes the projection from 3D space to image coordinates. The reprojection error of a point

is defined as the Euclidean distance between the predicted projection position p_i, i.e.

$$E_p(i) = \left\| \tilde{p}_i - p_i \right\|_2 \tag{15}$$

For a line segment, we project the predicted endpoints $\hat{A}_j$ and $\hat{B}_j$ onto the image plane, and then calculating the distance between the projected line segment and the observed line segment. To improve the robustness to outliers and depth anomalies, the above point-line reprojection errors are aggregated into point-line reprojection losses L^p_{proj} and L^l_{proj} through a robust loss ReproLoss with a dynamic threshold, and gradually introduced at a certain pace during training, so that 3D regression and uncertainty learning are the main focus in the early training stage, and geometric consistency constraints are gradually strengthened in the later stage. Combining all terms, the overall loss for PL2Map is expressed as follows:

$$L_{\text{PL2Map}} = L^p_{\text{3D}} + L^l_{\text{3D}} + L^p_{\text{unc}} + L^l_{\text{unc}} + L^p_{\text{proj}} + L^l_{\text{proj}} \tag{16}$$

During inference, the network exclusively outputs 2D-3D correspondences between points and line segments along with their uncertainty estimates, while the final 6-DoF pose is resolved by a backend geometric solver. This architecture, combining MoE-Attentional GNN intermediate representation, scene coordinate regression, and geometric reprojection constraints, achieves a balance between computational efficiency and robust scene structure perception. The hierarchical design provides high-quality teacher features for subsequent ESHD distillation by capturing richer multimodal geometric information.

4.4 Hierarchical distillation

Although the introduction of the MoE significantly enhances the model's ability to adapt to different visual patterns, the computational and storage overhead caused by the multi-expert structure is not conducive to deployment on resource constrained robot platforms. To address this issue, this paper designs the ESHD between the MoE-PL2Map teacher network and the compact student network. However, if the teacher and the student are only aligned at the final pose output layer, the student will have difficulty recovering the structural priors of the teacher in intermediate geometric representations and expert selection behaviors, and it is prone to the issue of "numerical proximity but structural deviation".

For this reason, the proposed ESHD framework decomposes knowledge distillation into three complementary pathways: feature-level, gate-level, and structure-level (Hint) distillation. This enables the student model to progressively approximate the teacher's structured decision-making process while maintaining compactness. Crucially, this partitioning strictly follows the geometric reasoning chain of the teacher network—comprising input encoding, MoE pattern decomposition, point-line message passing, and 3D coordinate regression—rather than applying uniform ℓ_2 distillation on equidistant backbone layers. Consequently, ESHD simultaneously constrains geometric outputs, routing behaviors, and multi-layer semantic structures across the reasoning pipeline. Unlike traditional "hierarchical distillation" that merely aligns features across select convolutional or FPN layers, our design offers dual advantages: 1) enabling the student model to approximate the teacher's input geometric distributions at shallow layers while reproducing structured GNN representations at deep layers; 2) explicitly preserving MoE's pattern decomposition capability through gating distillation. As a result, without increasing the inference overhead, the lightweight student model can get closer to the high-capacity MoE teacher model in terms of geometric accuracy, robustness in scene switching, and training stability.

Structurally, the teacher model adopts the complete MoE–PL2Map architecture, concatenating an Attentional GNN with a "self–cross–self–cross–self" topology on point-line geometric features, and learns structural preferences under different geometric modes through a MoE module with routers. On the premise of keeping the input and output interfaces consistent, the student model compresses the attention dimension and the number of GNN layers. It also removes all experts and routing structures and only retains single path feature propagation, thereby significantly reducing the number of parameters and the amount of computation. To inherit the expert selection mechanism of the teacher model without incurring the computational overhead of MoE, we attach a lightweight gating head to the GNN output layer of the student network. This gating head consists of two layers of 1×1 convolutions and ReLU activations, and it outputs gating logits with the same number of experts as the teacher. Operating exclusively during training, the branch approximates the teacher's gating distribution through temperature softening Jensen-Shannon divergence and entropy regularization, thereby preserving MoE's structured inductive bias while maintaining model compactness.

4.4.1 Feature-level distillation

Feature-level distillation directly aligns the 3D point-line output and geometric descriptors of teachers and students, ensuring that students can still inherit the teacher's physical geometry priors without reprojection loss. Let the feature of teachers and students in the integrated geometric coding space is denoted as f^T and f^S , respectively. The feature-level distillation loss is defined as a weighted combination of Euclidean distance and cosine similarity, i.e.

$$L_{\text{feat}} = \gamma \left\| f^S - f^T \right\|_2^2 + (1-\gamma)\left(1 - \frac{\langle f^S, f^T \rangle}{\|f^S\|_2 \, \|f^T\|_2}\right) \tag{17}$$

where $\gamma \in [0,1]$ is a balance coefficient used to balance the mean square error (MSE) term and cosine similarity, and $\gamma = 0.5$ in this paper. This loss encourages students to reproduce the teacher's discriminative feature distribution in a joint point-line geometric space.

4.4.2 Gate-level distillation

Gate-level distillation aligns the expert selection probability of MoE, transferring the routing behavior of teachers across different geometric patterns to students without explicit expert structures. Herein, we utilize the routing distribution of the teacher MoE as a "behavioral prior" to constrain the expert selection pattern of the student's lightweight gating head. Let the expert selection probability distributions generated by the teacher and student gate networks for the same input be denoted as $p^T = \{\alpha_i^T\}_{i=1}^N$ and $p^S = \{\alpha_i^S\}_{i=1}^N$, respectively. Then, routing preferences can be aligned by minimizing the JS divergence between the two distributions and applying entropy regularization. The gate-level distillation loss is defined as follows:

$$L_{\text{gate}} = \tfrac{1}{2} D_{\text{KL}}\big(p^T \,\|\, m\big) + \tfrac{1}{2} D_{\text{KL}}\big(p^S \,\|\, m\big), m = \tfrac{1}{2}(p^T + p^S), \tag{18}$$

In Eq. (18), L_{gate} leverages the teacher's gate distribution p^T as soft labels, explicitly constraining the student's gate head to reproduce the teacher's preference for using each expert at every frame. This enables the student to learn the teacher's implicit pattern segmentation capabilities across different structural regions even after removing the explicit MoE structure, thereby significantly enhancing the student's structural consistency across varying geometric areas. As a note, both MoE gating and gating distillation use room temperature Softmax (temperature $\tau = 1$),without the additional use of temperature annealing mechanism.

4.4.3 Hint-level distillation

Hint-level distillation selects four key stages: point-line input features + GNN output features, aligning the intermediate representations at two semantic depths: "geometric input" and "after geometric interaction". By aligning the hidden features of the teacher and the student at several key intermediate layers, the student can still maintain the same geometric perception and context modeling ability as the teacher even when the model structure is compressed. Let be the index set of these intermediate layers, and the features of teachers and students at several intermediate layers be h_k^T and h_k^S, respectively. The hierarchical hint distillation loss is defined as follows:

$$\begin{aligned} L_{\text{hint}} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \Bigg[& (1-\eta) \left\| h_k^S - h_k^T \right\|_2^2 \\ & + \eta \left(1 - \frac{\langle h_k^S, h_k^T \rangle}{\|h_k^S\|_2 \, \|h_k^T\|_2} \right) \Bigg] \end{aligned} \tag{19}$$

where $\eta \in [0,1]$ is the balance coefficient used to adjust preferences between Euclidean distance and cosine similarity, and $\eta = 0.8$ in this paper. This term encourages the student to reproduce the teacher's organization of geometric structures across multiple semantic levels, thereby preserving the teacher's structural priors even under a compressed model structure. Hint-level distillation complements supervision at the final output layer by enforcing alignment at several intermediate semantic depths, ensuring that the student maintains similar information flow and geometric inductive biases during feature transformations and fundamentally improving the representation quality and robustness of the lightweight model.

Combining features distillation, gate distillation, and hierarchical hint distillation, ESHD's overall distillation objective is as follows:

$$L_{\text{ESHD}} = \lambda_f L_{\text{feat}} + \lambda_g L_{\text{gate}} + \lambda_h L_{\text{hint}} \tag{20}$$

where λ_f, λ_g, and λ_h are hyperparameters for balancing the weights of each distillation term. In our case, $\lambda_f = 0.7$, $\lambda_g = 0.3$, and $\lambda_h = 0.5$ for all datasets. To describe the general training procedure, we express the optimization objective of the student network at the epoch as follows:

$$L_{\text{student}}(t) = w(t)\, L_{\text{ESHD}} \tag{21}$$

where $w(t)$ is the weight function that controls the overall distillation intensity, and expressed as follows:

$$w(t) = \begin{cases} w_{\text{init}}, & t < t_1 \\ w_{\text{init}} + \dfrac{w_{\text{final}} - w_{\text{init}}}{t_2 - t_1}\,(t - t_1) & t_1 \le t < t_2 \\ w_{\text{final}}, & t \ge t_2 \end{cases} \tag{22}$$

where w_{init} and w_{final} denote the distillation weights in the early and late stages of training, respectively; t_1 and t_2 are time windows that control the smooth transition of weights from weak to strong. Intuitively, the distillation intensity w_{init} is maintained at a low value during early training to weaken constraints and stabilize the student's geometric representations. In the mid-training phase, weight is gradually increased within the $[t_1, t_2]$range, enabling progressive alignment of the student network with the teacher's structured representations and expert selection behaviors. Finally, a high w_{final} value is applied in late training to intensify approximation of the teacher's outputs. Hence, Eq. (22) provides a unified scheduling description that facilitates scaling training strategies as needed for larger-scale or more complex scenarios, without compromising the consistency of the experimental configuration and implementation described in this paper.

Additionally, since the gating network and multiple expert branches in MoE-ESHD may produce large gradient norms in certain training stages, we apply global ℓ_2 norm clipping [40] to all learnable parameters to avoid gradient explosion and improve numerical stability. Let $\mathbf{g}$ denote the concatenated gradient vector for all parameters in the current mini-batch, then the clipped gradient $\mathbf{g}'$ is defined as follows:

$$\mathbf{g}' = \frac{\mathbf{g}}{\max\big(1, \|\mathbf{g}\|_2 / \delta\big)} \tag{23}$$

where δ is the clipping threshold, and $\delta = 5$ in this paper. This operation restricts the magnitude of the gradient while keeping its direction unchanged, which

can reduce the gradient fluctuations caused by the relatively sharp distribution of gating weights and improve the convergence stability of the MoE - ESHD training.

Through the above multi-level distillation constraints, the student model, despite having no explicit expert structure, maintains behavior in the structural subspace that is highly consistent with the teacher. As a result, it achieves stable and robust geometric representations under the inference-time efficiency constraints of humanoid robot deployment scenarios.

5 Experiments

In this section, we first evaluate the proposed algorithm on several public benchmarks and then validate its performance in real-world experiments on a humanoid robot. Unless otherwise specified, all experiments are conducted on a workstation equipped with an Intel i9–14900K CPU and 32 GB DDR5 RAM, running Ubuntu 20.04 LTS.

5.1 Simulation studies

To validate the performance of the proposed visual relocalization method under diverse scene conditions, we selected three representative datasets for benchmarking: 1) the small-scale indoor dataset "7-Scenes [19]", characterized by weak textures, severe occlusions, and significant illumination variations; 2) the large-scale outdoor dataset "Cambridge Landmarks [7]", featuring intense lighting contrasts and viewpoint changes; 3) the "Indoor-6 [51]" dataset, which captures semi-static scene evolution (including appearance drift and local structural changes) across different time periods.

All experiments were conducted using officially partitioned training/test splits, and input images were uniformly resized to a long edge of 1024 pixels while maintaining their original aspect ratios via bilinear interpolation. Evaluation metrics included: Median Pose Error (MPE), Recall@5cm, 5°, Absolute Trajectory Error (ATE) [52] in continuous sequences. Also, comparative analysis was performed against state-of-the-art methods spanning depth regression, geometric learning, and feature matching, including: PoseNet2 [41], VLocNet [42], DSAC++ [11], KFNet [44], DSM [45], DFAM [46], and HLoc [53].

5.1.1 Positioning error

Table 1 lists the MPE of different methods on the 7-Scenes sequence. It can be observed that the MoE's teacher model achieved the best or second-best performance in scenes such as Chess, Fire, and Heads, with an overall average error reaching 0.03 m / 0.87°. In the cases of weak texture and severe occlusion, the introduced point-line geometric constraints significantly enhance the robustness of pose estimation. As a result, the model's stability in these challenging situations was superior to that of depth regression methods [41, 42] and some geometric learning methods [45, 46]. In addition, while significantly compressing the model size, the student model still basically maintained a performance level similar to that of the teacher model. In most scenes, the accuracy of the student model was better than that of depth regression methods such as PoseNet2 [41] and VLocNet [42]. Therefore, it combined the advantages of lightweight design and high-precision.

Table 2 further compares the MPE of different methods on the Cambridge Landmarks sequences. We observe that the MoE model achieved localization accuracy comparable to geometric consistency methods like DSM [45] and HSC-Net [49] in scenes including

Table 1 Comparison of the MPE among different methods on the 7-Scenes sequence (m/°).

Method	Chess	Fire	Heads	Office	Pumpkin	RedKitchen	Stairs	Average
PoseNet2 [41]	0.13/4.48	0.27/11.3	0.17/13.0	0.19/5.55	0.26/4.75	0.23/5.35	0.35/12.4	0.23/8.12
VLocNet [42]	0.04/1.71	0.04/5.34	0.05/6.64	0.04/1.95	0.04/2.28	0.04/2.20	0.10/6.48	0.05/3.80
CamNet [43]	0.04/1.73	0.03/1.74	0.05/1.98	0.04/1.62	0.04/1.64	0.04/1.63	0.04/1.51	0.04/1.69
KFNet [44]	0.02/0.65	0.02/0.90	0.01/0.82	0.03/0.69	0.04/1.02	0.04/1.16	0.03/0.94	0.03/0.88
DSM [45]	0.02/0.68	0.02/0.80	0.01/0.80	0.03/0.78	0.04/1.11	0.03/1.11	0.04/1.16	0.03/0.92
DFAM [46]	0.02/0.64	0.02/0.73	0.01/1.07	0.03/0.75	0.04/0.91	0.03/1.03	0.04/1.14	0.03/0.89
CROSSFIRE [47]	0.01/0.4	0.05/1.9	0.03/2.3	0.05/1.6	0.03/0.8	0.02/0.8	0.12/1.9	0.04/1.1
Ours	0.02/0.65	0.02/0.78	0.01/0.73	0.03/0.76	0.04/0.94	0.04/1.22	0.06/1.02	0.03/0.87
Student	0.02/0.66	0.03/0.95	0.02/1.01	0.04/1.07	0.04/1.19	0.04/1.22	0.09/1.03	0.04/0.96

Table 2 Comparison of the MPE among different methods on the Cambridge Landmarks sequence (m/°).

Method	Great Court	Kings College	Old Hospital	Shop Facade	St M. Church	Average
PoseNet2 [41]	7.00/3.7	0.99/1.1	2.17/2.9	1.05/4.0	1.49/3.4	2.54/3.02
VLocNet [42]	NA	0.84/1.42	1.07/2.41	0.59/3.53	0.63/3.91	0.78/2.81
DSAC++ [11]	0.4/0.2	0.18/0.3	0.20/0.3	0.06/0.3	0.13/0.4	0.19/0.3
EMOSC [48]	1.35/0.70	0.25/0.39	0.34/0.75	0.65/0.68	0.26/0.56	0.57/0.62
HSC-Net [49]	0.28/0.2	0.18/0.3	0.19/0.3	0.06/0.3	0.09/0.3	0.16/0.3
DSM [45]	0.19/0.43	0.35/0.19	0.38/0.23	0.30/0.06	0.34/0.11	0.31/0.20
CCP-Net [50]	0.33/0.25	0.14/0.26	0.20/0.35	0.05/0.25	0.09/0.30	0.16/0.28
Ours	0.21/0.50	0.11/0.16	0.18/0.35	0.09/0.40	0.20/0.58	0.15/0.37
Student	0.27/0.63	0.13/0.21	0.26/0.43	0.10/0.40	0.34/0.99	0.22/0.53

Table 3 Comparison of the Recall@5cm,5° among different methods on the Indoor-6 sequence.

Method	scene1	scene2a	scene3	scene4a	scene5	scene6	Average
DSAC* [54]	18.7	28.0	19.7	60.8	10.6	44.3	30.4
SLD [51]	35.0	34.6	50.8	56.3	43.6	38.9	44.9
NBE+SLD [51]	38.4	–	53.0	–	40.0	50.5	45.5
Hloc-l1000 [53]	33.3	12.5	48.3	34.8	21.9	47.4	33.0
Hloc-l3000 [53]	48.1	17.1	61.9	39.2	31.1	59.1	42.8
SLD* [55]	68.5	62.6	76.2	77.2	57.8	78.0	70.1
Ours	58.5	66.2	63.2	83.5	42.2	59.1	62.1
Student	46.3	41.6	53.0	65.2	30.7	46.4	47.2

Kings College, Shop Facade, and Old Hospital. Particularly in environments with significant scale variations and strong illumination contrasts (e.g., Kings College), the point-line geometric constraints effectively mitigated pose drift, maintaining consistent accuracy. Notably, despite substantial parameter reduction, the student model preserved the teacher's performance characteristics, demonstrating robust accuracy in large-scale scenes without significant degradation due to model compression.

From Table 3, it can be seen that the average Recall@5cm, 5° of the MoE's teacher model reached 62.1%, second only to SLD* [55], and it significantly outperformed DSAC* [54], SLD [51], and HLoc [53] in all time-varying scenes. Even when faced with appearance drifts and local layout changes across time periods, the student model can still maintain a performance trend consistent with that of the teacher model, demonstrating good cross-time robustness. It is worth mentioning that in scenarios with drastic temporal changes (such as the same location under different weather or lighting conditions), the fluctuation range of the positioning error of the student model was comparable to that of the teacher model. This indicates that the lightweight model obtained through hierarchical distillation still has excellent robustness under time-varying conditions [51].

To further evaluate trajectory estimation stability, Fig. 3 compares the performance of Ours, CamNet [43], and CCP-Net [50] across: Chess(a) and Fire(b) sequences from 7-Scenes dataset, Kings College(c) and Shop Façade(d) sequences from Cambridge Landmarks dataset, and Scene4a(e) sequence from Indoor-6 dataset. The MoE's teacher model demonstrated near-perfect alignment with ground-truth trajectories in indoor scenarios (Chess, Fire), whereas CamNet exhibited significant drift (up to 0.35m) during abrupt lighting changes or partial occlusions. In outdoor environments (Kings College, Shop Façade), the MoE model maintained smooth trajectory estimation despite large parallax and strong illumination contrasts, while CCP-Net produced discontinuous trajectory segments under overexposure or extreme viewpoint changes. For the time-varying Scene4a sequence,

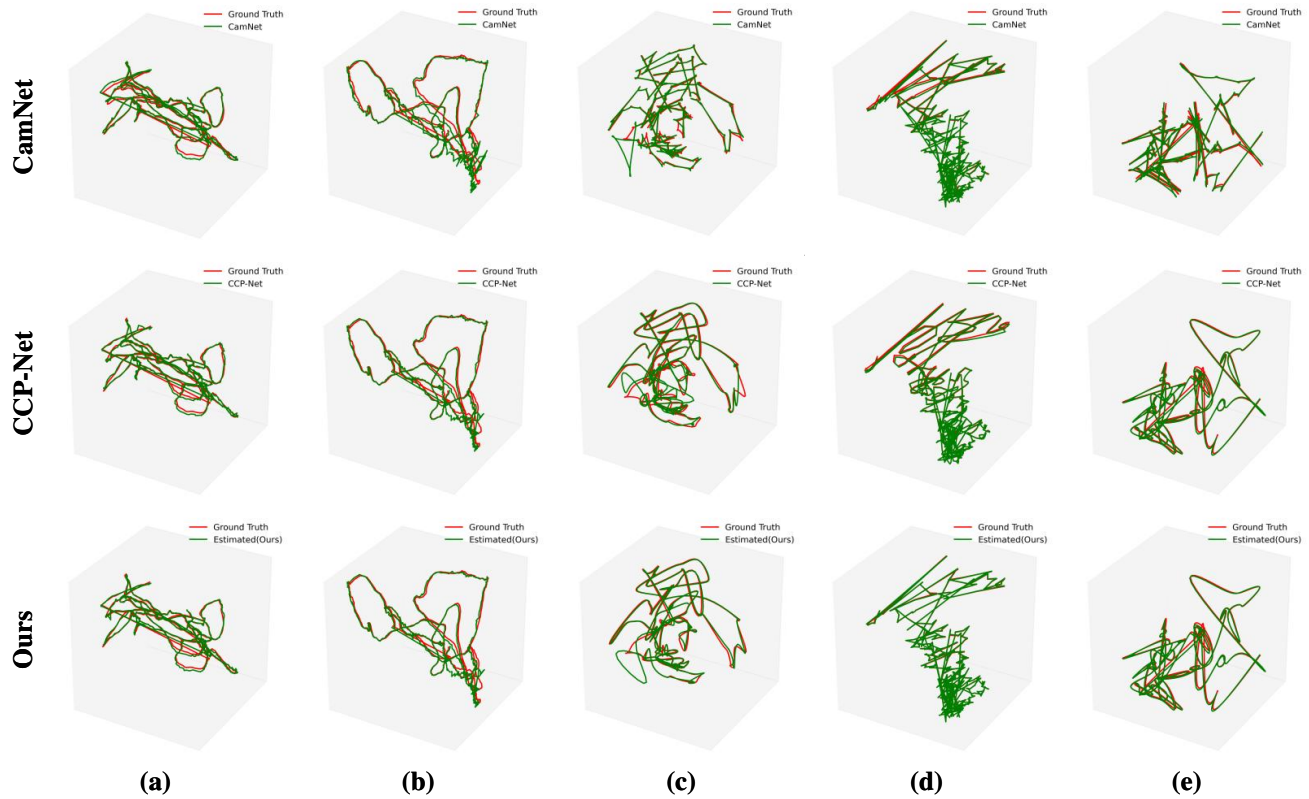


Fig. 3 Comparison of the generated trajectory among different algorithms.



Fig. 4 Attention distribution of experts on different sequences. Odd rows indicate input images, while different colors in even rows indicate different expert distributions.

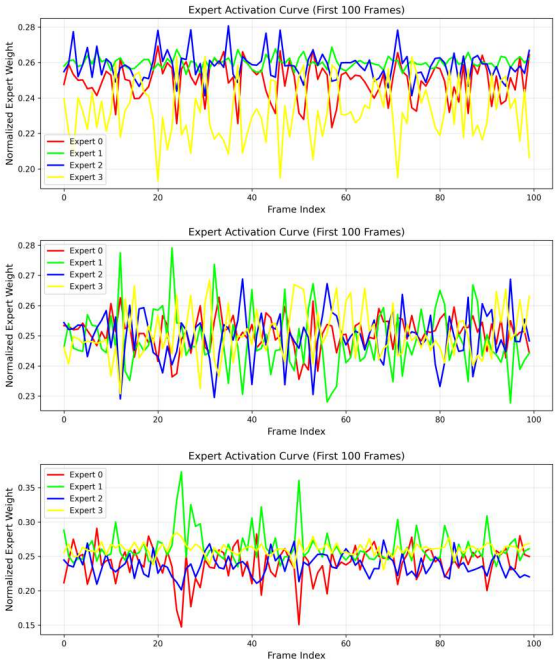


Fig. 5 Changes in the weights of experts on different sequences over time.

the MoE trajectory remained highly consistent with reference poses despite calibration errors, furniture displacement, and dynamic objects, showing negligible drift accumulation.

5.1.2 Expert visualization

To assess the effectiveness of MoE's dynamic expert mechanism during actual inference, we conducted a visualization analysis of expert distribution regions and weights across diverse scenarios.Fig. 4 illustrates the focal regions (i.e., expert attention distributions) of individual experts in representative indoor, outdoor, and time-varying scenes, while Fig. 5 quantitatively plots the temporal evolution curves of expert weights in corresponding environments. From these results, we could reasonably to conclude that: 1) Specialized expert allocation: each expert demonstrates clear functional specialization, where expert subnetworks biased toward point-based features were preferentially activated in texture-dense or corner-rich regions, while those favoring line-structured representations exhibited higher activation in scenes with well-defined geometric contours and dominant long straight lines; 2) Strong temporal consistency: the expert weights exhibited smooth temporal evolution across consecutive frames, demonstrating that expert selection maintains both continuity and stability under sequential input conditions; 3) Good scene adaptability: indoor scenes predominantly activate experts specializing in local textures and small-scale structures, while outdoor scenes require collaborative multi-expert configurations to handle scale variations and illumination differences. In time-varying conditions, expert switching undergoes gradual adaptation aligned with scene appearance transitions. This further reveals that the dynamic expert selection mechanism enables the model to adaptively adjust the structural representation according to the input features, thereby significantly enhancing the spatial consistency and temporal coherence.

5.1.3 Ablation experiment

To evaluate the individual contributions of point-line geometric constraints, MoE integration, and hierarchical distillation to system performance, we construct

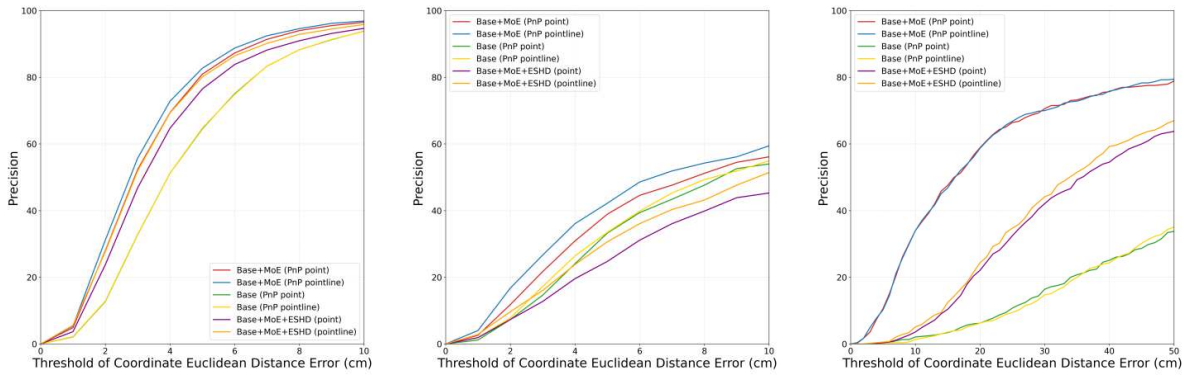


Fig. 6 Recall under different thresholds on the Indoor-6 sequence.

Table 4 Comparison of efficiency among different models.

Methods	Parameters count (M)	Model size (MB)	Inference time (ms)
Teacher (Ours)	16.49	69.3	19.5
Student (ESHD)	5.2	12.7	33.2
DSM [45]	25.1	100.4	29.7
Hloc [53]	–	> 500	~ 1000

ablation configurations across two dimensions (network architecture and pose estimation strategies) while maintaining identical training data, hyperparameters, and evaluation protocols. Architectures include: Baseline model "Base", Baseline integrated with MoE "Base+MoE (Teacher)", Teacher model compressed via ESHD distillation "Base+MoE+ESHD (Student)". Geometric configurations comprise: Point-feature-based PnP solver(PnP-point) and PnP with point-line reprojection constraints(PnP-pointline). Fig. 6 depicts Recall curves for these variants on three representative sequences: 7-Scenes/Office, Indoor-6/Scene5, and Cambridge/StMarys. Notably, higher curve positions at identical thresholds indicate stronger relocalization robustness.

From the perspective of geometric constraints, PnP-pointline consistently outperformed PnP-point, indicating that line features provide effective structural supplementation in pose solving. In the "Office" scene with abundant point features, the introduction of line constraints was mainly reflected in the improvement of the recall rate in the medium-to-high threshold regions and the reduction of the long-tail failure ratio. This shows that lines mainly play a role in enhancing the stability of the solution. In "Scene5" with more obvious weak textures and local occlusions, the curve of PnP-pointline rised more prominently in the strict threshold interval. This indicates that when point features degrade, line contours can still provide reliable constraints for direction and scale estimation, thereby improving geometric observability and solution stability. In the large-scale outdoor "StMary" scene, point-line constraints also bringed stable but relatively limited benefits. This implies that the performance bottleneck of such sequences is more restricted by feature representation and cross-scene generalization, rather than just by back-end geometric modeling.

Additionally, after the introduction of MoE, the recall curves of the three scenes shifted upward as a whole, and the improvement was more significant in the StMary scene. According to the visualization of expert routing, when texture information diminishes, the activation weights of geometric experts increased significantly. Under dynamic interference or local detail variations, texture experts demonstrate enhanced responsiveness. In sequences with substantial viewpoint changed and large depth spanned, collaborative reasoning involving multiple experts frequently occurs. This demonstrates that MoE employs sparse gating for conditional routing of perception modes (e.g., texture-dominant vs. structure-dominant), thereby improving scene coordinate regression consistency across viewpoint changes and distribution shifts.

Furthermore, comparative analysis of "Base+MoE +ESHD" versus "Base+MoE (Teacher)" reveals that, with approximately a 70% reduction in the number of parameters, the student model maintained comparable levels to the teacher model in terms of average localization error, cross-scene stability, and temporal trajectory smoothness. This indicates that the proposed ESHD strategy can effectively transfer the structural advantages of MoE to a compact network without incurring additional inference overhead, thereby achieving both lightweight design and high precision. From this, we can conclude that the point-line joint PnP provides robust geometric constraints, the MoE module enhances adaptability to complex visual patterns and scene distributions, while the ESHD strategy preserves these representational gains in lightweight networks. Their synergistic collaboration collectively supports the proposed framework's high-performance relocalization across diverse scenes.

To compare the efficiency of the model, Table 4 lists the MoE's teacher model, ESHD student model,

and multiple baselines in terms of parameter quantity, model footprint, and inference latency. We find that the ESHD model achieved approximately 1/8 the model size of DSM while maintaining comparable performance, highlighting the proposed framework's superior engineering deployability. This compactness makes it notably suitable for achieving real-time visual relocalization on humanoid robotic platforms.

5.2 Experimental testing

To evaluate the performance and effectiveness of the proposed method in real-world scenes, we deployed the MoE's teacher model and its ESHD student model on the "KUAVO-4pro" humanoid robot and conduct experiments in indoor and corridor environments. These environments include typical challenging conditions such as dynamic objects, weak textures, and strong illumination changes, as illustrated in Fig. 7.

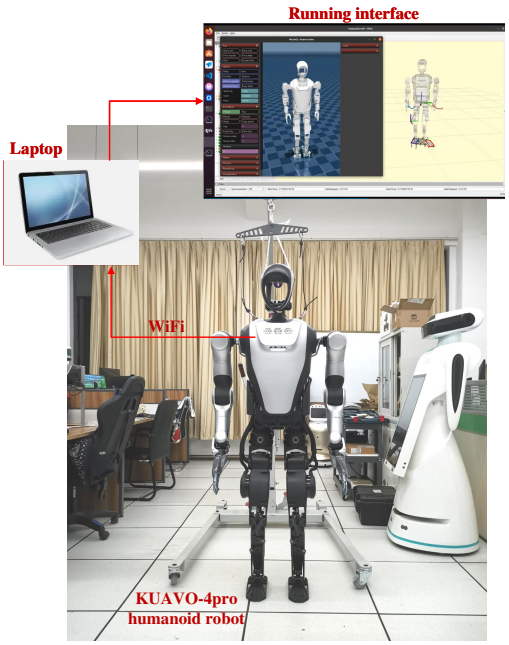
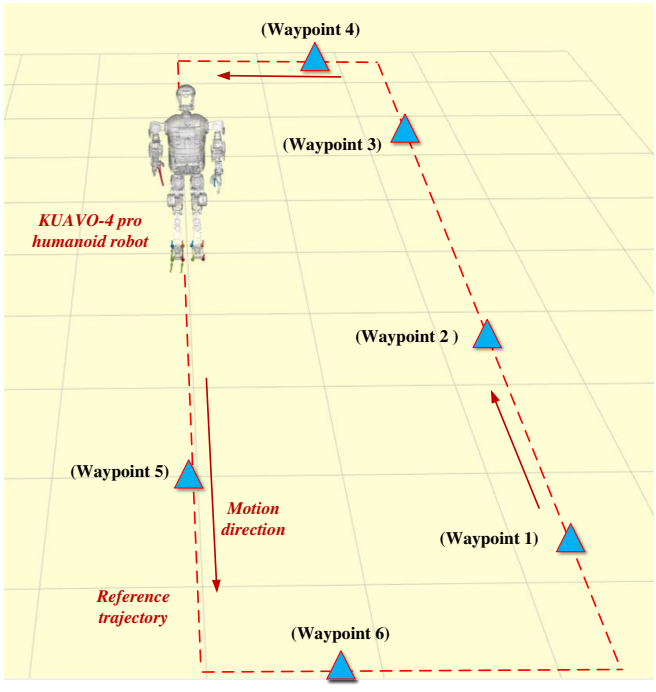


Fig. 7 Experimental scene and humanoid robot.

Fig. 8 depicts the expert response patterns of the MoE model at six distinct waypoints traversed by the robot in an indoor environment. It can be observed that the activation distributions among experts exhibited significant variations in response to environmental changes. When the robot entered low-texture or abruptly illuminated regions (Waypoints 3 and 5), geometric experts received substantially increased weights to compensate for uncertainty caused by texture degradation. Conversely, under dynamic object interference (Waypoint 2), texture-oriented experts show enhanced responses to maintain local detail continuity and estimation stability. This demonstrates that the MoE's teacher model dynamically adjusts expert weights according to input feature variations, achieving adaptive encoding across diverse visual conditions.

Fig. 9 further compares the predicted trajectory (green) against the ground-truth trajectory (red) during a full navigation cycle of the humanoid robot in the indoor environment. Meanwhile, we also constructed a 3D dense point cloud map of the scene to better demonstrate the localization effects. It can be observed that at Waypoint 1 (static and stable) and Waypoint 2 (low-texture), the predicted trajectory of the student model almost completely coincided with the reference trajectory. At Waypoint 2 (dynamic interference), despite the presence of moving objects, the generated trajectory can still fit the reference trajectory smoothly and continuously. At Waypoint 5 (strong illumination change), the estimated trajectory still maintained good consistency with the reference trajectory. Evidently,



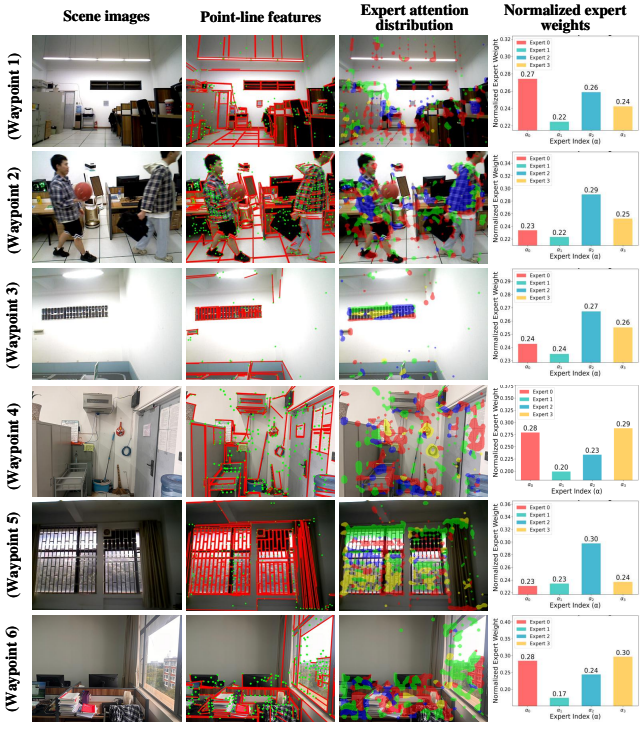


Fig. 8 Visualization results of various experts at different waypoints of the teacher model in an indoor scene.

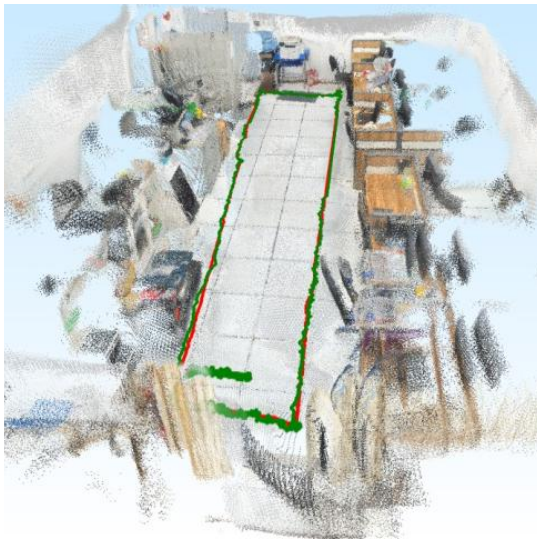


Fig. 9 The trajectory generated by student model deployed on the humanoid robot in indoor scene.

the student model inherited the teacher model's temporal stability and spatial consistency across different scenes, highlighting the advantages of structural representation fidelity and deployment efficiency brought by hierarchical distillation.

To validate the performance of our method in outdoor scenes, we conducted experimental tests in a 27 m × 20 m corridor. This scene represents a typical "structural repetition + texturelessness + dynamic interference" scenario prone to scale drift and geometric degeneracy. Fig. 10 compares the ESHD model's generated trajectory (green) against the ground-truth trajectory (red), while Fig. 11 further visualizes scene characteristics and extracted features at key waypoints. Also, a complementary 3D dense point cloud map was reconstructed for context. We find that, despite challenges posed by weak textures, dynamic disturbances, and structural repetition, the ESHD trajectory maintained consistent alignment with reference poses throughout the humanoid robot's navigation, exhibiting no significant deviation or distortion. This demonstrates that the proposed distillation mechanism effectively preserves the teacher model's spatial geometric representation capabilities, enabling the lightweight student model to adapt to complex environments.

Fig. 12 quantitatively describes the curve of the ATE over time in the corridor. It can be observed that throughout the entire process, the ATE of the ESHD model always stayed within a small fluctuation range. Even though there were some deviations in the weakly - textured areas, there was no obvious error accumulation. Even though local texture degradation can cause short-term accuracy disturbances, the error was always kept within an acceptable range, and

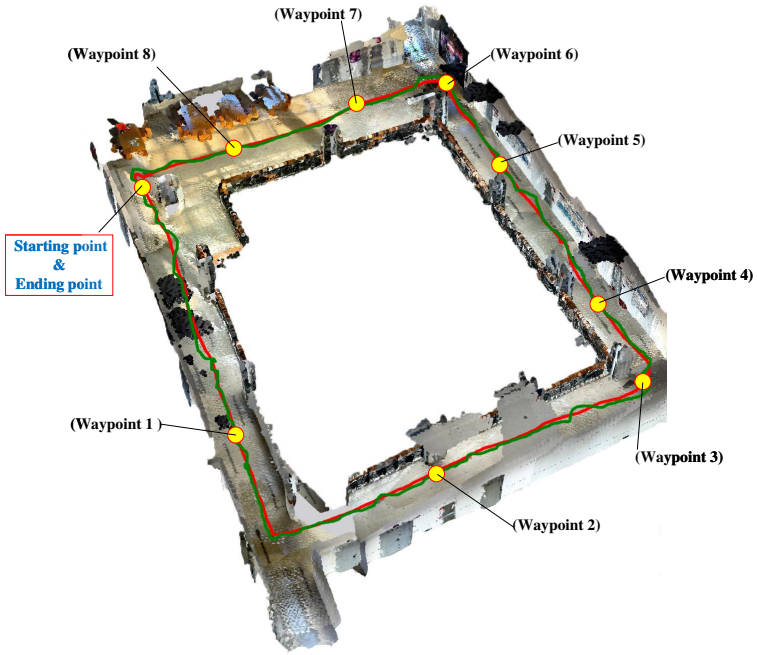


Fig. 10 The trajectory generated by the student model in the corridor.

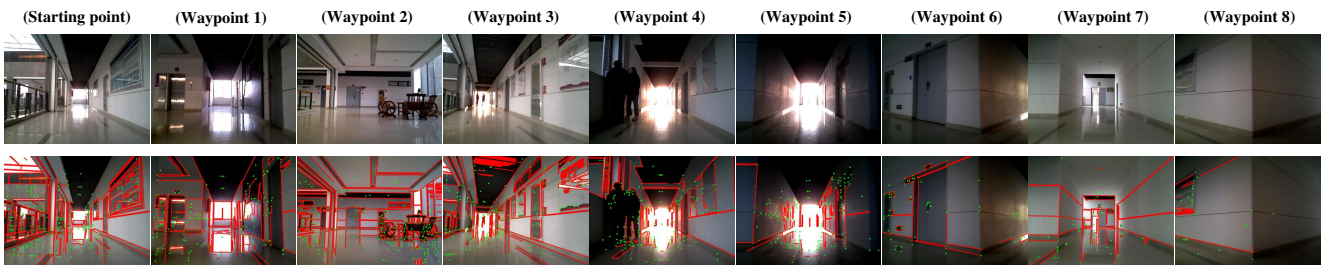


Fig. 11 Scene images at different waypoints (top row) and extracted features (bottom row) in corridor.

Table 5 The performance of algorithms in real scenes.

Scene	Average ATE (cm)	Runtime (s)	Max ATE (cm)	Std ATE (cm)	Recall@5cm,5° (%)
Indoor	2.1	45.6	6.4	1.0	97.2
Corridor	3.4	62.5	12.3	1.7	93.1

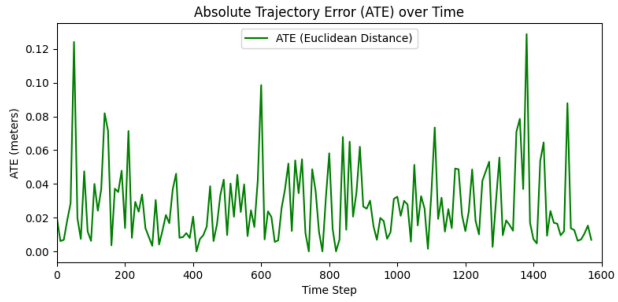


Fig. 12 ATE curve of the ESHD model in corridor.

the predicted trajectory can maintain a stable spatial trend. This indicates that the ESHD model not merely performs robustly in indoor scenarios but can also maintain long-term consistent localization ability in typical geometrically degenerate environments such as corridors.

From Table 5, it can be observed that in real indoor environments, the ESHD student model achieved an average ATE of approximately 2.1 cm, with a maximum error of 6.4 cm and a standard deviation of around 1.0 cm. The Recall@5cm,5° metric reached 97.2%. Error fluctuations remained consistently within a small centimeter-scale range over time, with only a few frames approaching the threshold boundary. Combined with the results in Fig. 6, it can be concluded that the student model provided stable and reliable localization results even under conditions of self-occlusion and localized weak textures. In corridor scene with more pronounced geometric degeneracy, characterized by increased long linear structures and repetitive patterns, the average ATE rised to 2.9 cm, with a maximum error of approximately 12.3 cm and a standard deviation of 1.7 cm. The Recall@5cm,5° metric slightly decreased to 93.1%. Compared to indoor environments, both error magnitude and fluctuations increase in corridors, but the overall error remains within a low range. The ATE curve exhibited a pattern of "localized minor spikes followed by rapid recovery", without sustained drift or error bursts. This demonstrates the student model's strong localization capability in large-scale scenes with rapid viewpoint changes and structural degradation.

Overall, the experimental results demonstrate that, despite substantially reduced parameters and computational overhead, the ESHD student model maintains centimeter-level accuracy during real-world deployment on the humanoid robot, while keeping Recall@5cm,5° consistently high across different scenarios. Moreover, the discrepancies between indoor and corridor scenarios primarily manifest in extreme frames and error fluctuations rather than catastrophic accuracy collapse, indicating that the hierarchical distillation with expert selection awareness effectively preserves the teacher model's geometric representation and robustness. This enables the lightweight student model to meet the comprehensive requirements of real-time performance and stability for long-term online relocalization in humanoid robots.

6 Conclusion and future work

This paper proposed an adaptive visual relocalization framework integrating MoE with ESHD, addressing the challenge of simultaneously achieving accuracy, stability, and real-time performance in humanoid robot relocalization within complex environments. The framework employed SuperPoint and DeepLSD for point-line feature extraction at the frontend of the constructed PL2Map scene coordinate regression network. An MoE-Attentional GNN module was inserted into PL2Map's intermediate geometric feature space to adaptively model diverse geometric patterns, enabling robust relocalization through point-line 3D regression and geometric reprojection losses. Notably, the ESHD hierarchical distillation mechanism jointly constrained feature-level, routing-level, and intermediate Hint-level representations, successfully transferring the teacher model's structured reasoning capabilities to a lightweight student model. This achieved significant model compression (≈70% parameter reduction) and inference acceleration (33.2 ms with the student model) while maintaining comparable accuracy and cross-scene consistency to the teacher. Experimental results on 7-Scenes, Cambridge Landmarks, Indoor-6 sequences, and the KUAVO-4pro humanoid robot demonstrate that the MoE-ESHD framework effectively balanced accuracy, robustness, and efficiency, providing a theoretically sound and practically valuable visual relocalization solution for resource-constrained humanoid robots.

Future work will focus on further enhancing system accuracy and generalization capabilities through two complementary directions. First, we intend to integrate advanced temporal modeling mechanisms (such as lightweight Transformers or Recurrent units) to explicitly exploit short-term kinematic constraints generated during humanoid gait cycles and upper-limb manipulations. This will improve suppression of pose drift during continuous operation by leveraging motion periodicity. Second, while ESHD has achieved significant model compression, we plan to investigate neural architecture search (NAS)-based automated optimization to realize ultra-low-power, ultra-low-latency inference on embedded processors (e.g., FPGA/ASIC), maintaining distillation fidelity. Additionally, we plan to extend the framework to multimodal relocalization by fusing inertial measurement units (IMU) or LiDAR, enabling robust operation in extreme scenarios (e.g., low-light or severe occlusion) where monocular vision

fails completely-a critical capability for humanoid autonomy in complex environments.

Funding. This work was supported by the National Natural Science Foundation of China under Grants 62463033 and 62063036, the "Xingdian Talent Support Program" Youth Talent Special Project of Yunnan Province under Grant 01000208019916008, and the Basic Research Program of Yunnan Province - Key Project under Grant 202501AS070014.

Data Availability. The datasets used in our experiments are available in https://www.microsoft.com/en-us/research/project/rgb-d-dataset-7-scenes/, http://mi.eng.cam.ac.uk/projects/relocalisation/#dataset, https://github.com/microsoft/SceneLandmarkLocalization, and https://drive.google.com/drive/folders/1w7Adnd6MXmNOacT072JnQ6emHUeLrD71.

Declarations.

Ethics approval. This study did not involve new data collection from human participants or animals.

Conflicts of interest. The authors declare that they have no conflicts of interest.

References

[1] Kumar, D., Muhammad, N.: A survey on localization for autonomous vehicles. IEEE Access **11**, 115865–115883 (2023) https://doi.org/10.1109/ACCESS.2023.3326069

[2] Stranner, M., Arth, C., Schmalstieg, D., Fleck, P.: A high-precision localization device for outdoor augmented reality. In: 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), pp. 37–41 (2019). https://doi.org/10.1109/ISMAR-Adjunct.2019.00025

[3] Li, J., Ding, W., Liu, G., Hua, C., Cai, D., Sun, J.: Excavator pose estimation under occlusion: a coordinate classification approach enhanced by Kalman filtering. The Journal of Supercomputing **81**(16), 1526 (2025) https://doi.org/10.1007/s11227-025-07989-y

[4] Liu, T., Luo, J.: Semantic-aware dynamic scene tracking and reconstruction for mobile robots leveraging visual SLAM. The Journal of Supercomputing **81**(7), 830 (2025) https://doi.org/10.1007/s11227-025-07343-2

[5] Lu, R., Zhu, Z., Fu, S., Chen, S., Wang, T., Yan, C., Xu, F.: Self-supervised camera relocalization with hierarchical fern encoding. IEEE Transactions on Instrumentation and Measurement **73**, 1–12 (2024) https://doi.org/10.1109/TIM.2023.3347802

[6] Sattler, T., Leibe, B., Kobbelt, L.: Efficient & effective prioritized matching for large-scale image-based localization. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(9), 1744–1756 (2017) https://doi.org/10.1109/TPAMI.2016.2611662

[7] Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2938–2946 (2015). https://doi.org/10.1109/ICCV.2015.336

[8] Sattler, T., Zhou, Q., Pollefeys, M., Leal-Taixé, L.: Understanding the limitations of cnn-based absolute camera pose regression. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3297–3307 (2019). https://doi.org/10.1109/CVPR.2019.00342

[9] Zhou, Q., Sattler, T., Pollefeys, M., Leal-Taixé, L.: To learn or not to learn: Visual localization from essential matrices. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 3319–3326 (2020). https://doi.org/10.1109/ICRA40945.2020.9196607

[10] Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C.: Dsac — differentiable ransac for camera localization. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2492–2500 (2017). https://doi.org/10.1109/CVPR.2017.267

[11] Brachmann, E., Rother, C.: Learning less is more - 6d camera localization via 3d surface regression. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4654–4662 (2018). https://doi.org/10.1109/CVPR.2018.00489

[12] Glocker, B., Shotton, J., Criminisi, A., Izadi, S.: Real-time rgb-d camera relocalization via randomized ferns for keyframe encoding. IEEE Transactions on Visualization and Computer Graphics **21**(5), 571–583 (2015) https://doi.org/10.1109/TVCG.2014.2360403

[13] Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer (2017). https://arxiv.org/abs/1701.06538

[14] Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network (2015). https://arxiv.org/abs/1503.02531

[15] Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: FitNets: Hints for Thin Deep Nets (2015). https://arxiv.org/abs/1412.6550

[16] Pumarola, A., Vakhitov, A., Agudo, A., Sanfeliu, A., Moreno-Noguer, F.: Pl-slam: Real-time monocular visual slam with points and lines. In:

2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 4503–4508 (2017). https://doi.org/10.1109/ICRA.2017.7989522

[17] Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4937–4946 (2020). https://doi.org/10.1109/CVPR42600.2020.00499

[18] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5297–5307 (2016). https://doi.org/10.1109/CVPR.2016.572

[19] Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene coordinate regression forests for camera relocalization in rgb-d images. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2930–2937 (2013). https://doi.org/10.1109/CVPR.2013.377

[20] Zhang, M., Zhou, Z.: Speed-accuracy trade-off in lightweight-based hand pose estimation. The Journal of Supercomputing **81**(12), 1212 (2025) https://doi.org/10.1007/s11227-025-07699-5

[21] Walch, F., Hazirbas, C., Leal-Taixé, L., Sattler, T., Hilsenbeck, S., Cremers, D.: Image-based localization using LSTMs for structured feature correlation (2017). https://arxiv.org/abs/1611.07890

[22] Valentin, J., Nießner, M., Shotton, J., Fitzgibbon, A., Izadi, S., Torr, P.: Exploiting uncertainty in regression forests for accurate camera relocalization. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4400–4408 (2015). https://doi.org/10.1109/CVPR.2015.7299069

[23] Cavallari, T., Golodetz, S., Lord, N.A., Valentin, J., Prisacariu, V.A., Stefano, L.D., Torr, P.H.S.: Real-time rgb-d camera pose estimation in novel scenes using a relocalisation cascade. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(10), 2465–2477 (2020) https://doi.org/10.1109/TPAMI.2019.2915068

[24] Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A.: InLoc: Indoor Visual Localization with Dense Matching and View Synthesis (2018). https://arxiv.org/abs/1803.10368

[25] Zhou, Q., Agostinho, S., Ošep, A., Leal-Taixé, L.: Is Geometry Enough for Matching in Visual Localization?, pp. 407–425 (2022). https://doi.org/10.1007/978-3-031-20080-9_24

[26] DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: Self-Supervised Interest Point Detection and Description (2018). https://arxiv.org/abs/1712.07629

[27] Fedus, W., Zoph, B., Shazeer, N.: Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity (2022). https://arxiv.org/abs/2101.03961

[28] Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Pinto, A.S., Keysers, D., Houlsby, N.: Scaling Vision with Sparse Mixture of Experts (2021). https://arxiv.org/abs/2106.05974

[29] Pavlitska, S., Fan, H., Ditschuneit, K., Zöllner, J.M.: Robust Experts: the Effect of Adversarial Training on CNNs with Sparse Mixture-of-Experts Layers (2025). https://arxiv.org/abs/2509.05086

[30] Gao, Y., Lan, W., Zhao, H., Huang, L., Liu, S., Guo, Q.: AnyExperts: On-Demand Expert Allocation for Multimodal Language Models with Mixture of Expert (2025). https://arxiv.org/abs/2511.18314

[31] Park, W., Kim, D., Lu, Y., Cho, M.: Relational Knowledge Distillation (2019). https://arxiv.org/abs/1904.05068

[32] Tung, F., Mori, G.: Similarity-Preserving Knowledge Distillation (2019). https://arxiv.org/abs/1907.09682

[33] Moslemi, A., Briskina, A., Dang, Z., Li, J.: A survey on knowledge distillation: Recent advancements. Machine Learning with Applications **18**, 100605 (2024) https://doi.org/10.1016/j.mlwa.2024.100605

[34] Zhang, H., Zhang, S.: A cascade knowledge distillation framework with three-stage training for dynamic knowledge graph embedding. The Journal of Supercomputing **82**(1), 29 (2025) https://doi.org/10.1007/s11227-025-08166-x

[35] Qian, Y., Wang, X., Sun, F., Pan, L.: Compressing transfer: Mutual learning- empowered knowledge distillation for temporal knowledge graph reasoning. IEEE Transactions on Neural Networks and Learning Systems **36**(8), 15309–15321 (2025) https://doi.org/10.1109/TNNLS.2025.3525699

[36] Feng, K., Sun, Z., Fang, H., Yang, J., Liu, W., Ong, Y.-S.: Routing Distilled Knowledge via Mixture of LoRA Experts for Large Language Model based Bundle Generation (2025). https://arxiv.org/abs/2508.17250

[37] Sadeghi, M., Rostami, M., Sadeghnejad, S.: Optimization dynamic stability and energy efficiency in human-like bipedal robot over a full gait cycle. In: Proceedings of the 32nd National and 10th International Conference on Biomedical

Engineering of Iran, Tehran, Iran (2025). Available at: https://tstaconf.ir/fa/Home/Article/4926b592-03f5-43f4-b490-cfc9e32f90ab

[38] Xu, J., Huang, Y., Pei, B., Hou, J., Li, Q., Chen, G., Zhang, Y., Feng, R., Xie, W.: EgoExoGen: Ego-centric Video Prediction by Watching Exo-centric Videos (2025). https://arxiv.org/abs/2504.11732

[39] Pautrat, R., Barath, D., Larsson, V., Oswald, M.R., Pollefeys, M.: DeepLSD: Line Segment Detection and Refinement with Deep Image Gradients (2023). https://arxiv.org/abs/2212.07766

[40] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library (2019). https://arxiv.org/abs/1912.01703

[41] Kendall, A., Cipolla, R.: Geometric loss functions for camera pose regression with deep learning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6555–6564 (2017). https://doi.org/10.1109/CVPR.2017.694

[42] Valada, A., Radwan, N., Burgard, W.: Deep auxiliary learning for visual localization and odometry. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 6939–6946 (2018). https://doi.org/10.1109/ICRA.2018.8462979

[43] Ding, M., Wang, Z., Sun, J., Shi, J., Luo, P.: Camnet: Coarse-to-fine retrieval for camera re-localization. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2871–2880 (2019). https://doi.org/10.1109/ICCV.2019.00296

[44] Zhou, L., Luo, Z., Shen, T., Zhang, J., Zhen, M., Yao, Y., Fang, T., Quan, L.: Kfnet: Learning temporal camera relocalization using kalman filtering. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4918–4927 (2020). https://doi.org/10.1109/CVPR42600.2020.00497

[45] Tang, S., Tang, C., Huang, R., Zhu, S., Tan, P.: Learning Camera Localization via Dense Scene Matching (2021). https://arxiv.org/abs/2103.16792

[46] Xie, T., Dai, K., Wang, K., Li, R., Wang, J., Tang, X., Zhao, L.: A deep feature aggregation network for accurate indoor camera localization. IEEE Robotics and Automation Letters **7**(2), 3687–3694 (2022) https://doi.org/10.1109/LRA.2022.3146946

[47] Moreau, A., Piasco, N., Bennehar, M., Tsishkou, D., Stanciulescu, B., La Fortelle, A.: Crossfire: Camera relocalization on self-supervised features from an implicit representation. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 252–262 (2023). https://doi.org/10.1109/ICCV51070.2023.00030

[48] Duong, N.-D., Soladié, C., Kacete, A., Richard, P.-Y., Royan, J.: Efficient multi-output scene coordinate prediction for fast and accurate camera relocalization from a single rgb image. Computer Vision and Image Understanding **190**, 102850 (2020) https://doi.org/10.1016/j.cviu.2019.102850

[49] Li, X., Wang, S., Zhao, Y., Verbeek, J., Kannala, J.: Hierarchical Scene Coordinate Classification and Regression for Visual Localization (2020). https://arxiv.org/abs/1909.06216

[50] Ruan, J., He, L., Guan, Y., Zhang, H.: Combining scene coordinate regression and absolute pose regression for visual relocalization. In: 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 11749–11755 (2023). https://doi.org/10.1109/ICRA48891.2023.10160317

[51] Do, T., Miksik, O., DeGol, J., Park, H.S., Sinha, S.N.: Learning to detect scene landmarks for camera localization. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11122–11132 (2022). https://doi.org/10.1109/CVPR52688.2022.01085

[52] Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 573–580 (2012). https://doi.org/10.1109/IROS.2012.6385773

[53] Sarlin, P.-E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12708–12717 (2019). https://doi.org/10.1109/CVPR.2019.01300

[54] Brachmann, E., Rother, C.: Visual camera relocalization from rgb and rgb-d images using dsac. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 5847–5865 (2022) https://doi.org/10.1109/TPAMI.2021.3070754

[55] Do, T., Sinha, S.N.: Improved scene landmark detection for camera localization. In: 2024 International Conference on 3D Vision (3DV), pp. 975–984 (2024). https://doi.org/10.1109/3DV62453.2024.00069