2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SLoFT: End-to-End Semantic Localization with Floorplan and Transformer

Chaerin Min[1,2], Hongsheng Yu[1], Fengtao Fan[1], Srinath Sridhar[2], Qiuxuan Wu[1], Chao Guo[1]

***Abstract*— Visual localization is critical for AR navigation, AI-driven audio guidance, and mobile robot localization. However, traditional SLAM methods that rely on pre-built 3D maps suffer from high costs, privacy concerns, and sensitivity to environmental changes. Recent floorplan-based localization methods attempt to address these challenges by using 2D floorplans, eliminating the need for 3D map construction. Still, existing approaches are often impractical for real-world applications, as they are limited to specific layouts and fail to generalize beyond their training domains. We propose a novel approach that learns to semantically match visual cues from a camera image to a floorplan image rich in semantic details, inspired by human ability to directly localize oneself using a complex floorplan image. To achieve this, we train a single, unified model on a diverse dataset of 1.2M images and 740K floorplans that we curated, which includes a new collection of semantically-rich, real-world floorplans. This allows our model to generalize effectively to previously unseen areas and implies generalization potentials to unseen buildings. Without making assumptions about camera poses or floorplan structures, our end-to-end model outperforms existing methods and enables variations like floorplan rotations, lighting changes, and different camera intrinsics.**

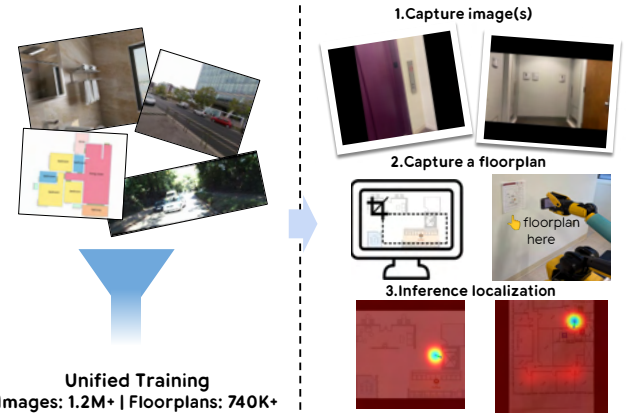

Fig. 1: We introduce a model that allows localization from 2D floorplan images. By training on diverse datasets, our model learns to interpret both screenshots and photographs, enabling effective localization across different environments.

I. Introduction

Camera localization is an important part of AR navigation, audio AI navigation for individuals with low vision, and mobile robots. For such goals, the essential task is to determine the device's 2D position and yaw (x, y, yaw) relative to a pre-built map, an important first step for providing any meaningful navigation instructions. Traditionally, building this map required comprehensive scanning of an area, capturing image sequences from every viewpoint to create a detailed database of visual features [1]–[4].

However, this approach suffers from several shortcomings. First, the initial cost and effort required to comprehensively scan every new building and street are high. This intensive setup process makes the approach difficult to scale across numerous different buildings and streets. Second, the process raises significant privacy concerns, particularly indoors, as capturing the detailed interiors of private spaces or individuals is often unavoidable, necessitating complex license procedures and data anonymization processes. Third, environments are highly dynamic, with frequent changes like shifting furniture or temporary objects, making the map quickly obsolete. This demands constant re-scans, which carry the same costs and privacy hurdles as the initial setup.

We propose a novel approach that circumvents these challenges by using 2D floorplans as a lightweight and robust alternative to 3D visual maps. First, this method reduces costs and complexity, as floorplans are often readily available from building directories or websites, removing the need for 3D map building. Second, by relying on an abstract layout instead of raw imagery, our approach inherently preserves privacy, as it does not capture or store any sensitive visual data. Third, because floorplans represent semantically essential structure of an environment, they are naturally robust to the dynamic local changes and temporary objects that make visual maps quickly outdated. In summary, our goal is to design an estimator that can effectively localize a camera within these floorplans, offering a practical, accessible, and scalable localization solution for AR and AI audio navigation.

To achieve this goal, the challenge is to bridge the gap between two modalities, floorplan and image. This has often been inherently difficult due to limited context in the floorplan. In this paper, our model learns to match semantic cues between these two modalities by treating the floorplan as a raw, unconstrained image with rich details rather than a simplified vector map. To train such a system, we curated a large-scale dataset of 1.2 million images and 740,000 floorplans. We then developed an end-to-end model, consisting of a dual encoder architecture and a fusion model to process both modalities and predict the camera's 3-DoF pose probability from a single image or optionally a sequence of images.

Our approach marks a significant departure from previous works that often required heavily simplified (e.g., walls-only) floorplans [5]–[9], imposed strict constraints on the

[1]Google; This work was performed when Chaerin was at Google. {hongshengyu, ffan, qiuxuanwu, chaoguo}@google.com
[2]Brown University. {chaerin_min, srinath}@brown.edu

camera orientation and height [5], [10], assumed known relative poses in image sequences [7], [11], or relied on floorplans encoded into predefined semantic channels [11]. Furthermore, unlike concurrent methods [8], [9] that train on specific datasets or domains (e.g., indoor or outdoor only), our model is trained on diverse domains of datasets in a unified manner. Consequently, our key contributions are summarized as follows:

- We propose an end-to-end model that leverages semantic cues from raw floorplans. This design removes the need for costly 3D map building or strict floorplan constraints, offering a practial localization solution, robust to dynamic changes and privacy-preserving.
- We demonstrate that a single model, trained on a unified dataset, generalizes across diverse domains, outperforming baselines trained on specific datasets.
- Our approach enables localization in novel environments at 13.3 FPS without retraining. It remains robust for lighting changes and enables diverse inputs, without additional test-time optimization process.

II. Related Works

Visual localization with 3D maps Traditional visual localization relies on prior built 3D scene representation of the venue to localize and one common approach is to use point clouds generated from Structure-from-Motion (SfM) or SLAM [12], [13]. Theses methods typically operate by matching local features [3] from a query image to the 3D map to determine the camera pose [1], [2]. In order to ensure long time operation success of these approaches, recent research within this direction focus on exploring alternative scene representation format [14], [15], incorporating semantic and object-level information to augment standard point cloud map [16], [17], improve robustness on short-term dynamic moving objects [18], [19] and perform change detection to handle long-term scene dynamics [16], [20], [21]. Alternative direction such as image retrieval and coordinate-based rendering have also been employed to find locations from a large image database [4], [22]. While accurate, this paradigm suffers from critical drawbacks, including the high cost of map creation and maintenance, privacy concerns, and sensitivity to dynamic environments.

Localization with floorplan To overcome the challenges of localization against prior-built 3D scene representation, there exist active research to achieve the same by using floorplan map. Vigor [23], which utilized satellite imagery, was limited to outdoor environments. Early methods for indoor localization from floorplans often relied on strong assumptions, such as an upright camera orientation [5], [10], [24], a known camera height [25], or explicitly a panoramic image [6]. These constraints limit their applicability for handheld or wearable devices like mobile phones or AR devices. Later research relaxed some of these constraints, but still had assumptions. For instance, F3Loc and its successor [7], [26] and OrienterNet [11] require known relative poses for image sequences. On the other hand [8] leverage room style knowledge and [9] uses semantics information on the floorplan edges to further improve accuracy. All these aforementioned approaches treat 2D maps as a collection of geometric primitives or leveraging specific domain knowledge on the floorplan style, neglecting rich semantic information like text and symbols that are generally available within the common floorplan maps. This necessitates a tedious map-parsing step and limits scalability. In contrast, our approach minimize assumptions. We use the raw floorplan image directly, requiring no prior information about camera parameters or map structure, thus offering a more general and practical solution.

III. Method

A. Problem definition and notation

Our goal is to estimate the 3-DoF pose of a camera, defined by its 2D position and yaw within a given floorplan. The inputs to our system are a query image $I \in \mathbb{R}^{H\times W\times 3}$ and a floorplan image $F \in \mathbb{R}^{H'\times W'\times 3}$. The estimated position $p = (x, y)$ is then defined within the pixel coordinate space $\mathbf{x} \in \{1, ..., H'\} \times \{1, ..., W'\}$ of the floorplan F. We formulate this as a learning task where our model, SLoFT, learns a function f that maps the two inputs to a localization result (x, y, θ). Instead of directly regressing the localization result, our model predicts the camera's position as a probability distribution across the floorplan, represented by a heatmap $H \in \mathbb{R}^{H'\times W'}$. The final 2D position is then inferred from this heatmap, such that $p = \text{argmax}_{\mathbf{x}} H(\mathbf{x})$. In addition to the position heatmap, the model predicts the yaw angle θ and an associated confidence score σ. As a result, SLoFT is a function that performs the following mapping:

$$(H, \theta, \sigma) = f(I, F) \tag{1}$$

B. Dual encoder

A key aspect of our approach is to process floorplans directly in their most common, scalable, and accessible format – images, without tedious preprocess steps that convert floorplans into structured graphs. Given this image-based representation for both the real-world query image and the abstract floorplan, our model designs a dual-encoder architecture. The goal to bridge the modality gap by mapping both inputs into a shared, high-dimensional semantic space where their features can be meaningfully compared. We utilize a Vision Transformer (ViT) architecture [27], [28], with its weights initialized from a model pretrained from large image datasets. The decision of using pretrained ViT encoders is validated in our ablation study.

C. Fusion module

From the encoders, we obtain two sets of feature tokens: $t_I \in \mathbb{R}^{K_I\times C}$ from the image encoder and $t_F \in \mathbb{R}^{K_F\times C}$ from the floorplan encoder, where K and C denote the number of tokens and the embedding dimension, respectively. The fusion module's role is to integrate these feature sets: identifying essential cues in the query image and matching them to the floorplan for a prediction of (x, y, yaw). To achieve this sparse goal, a set of K_q learnable queries [29],

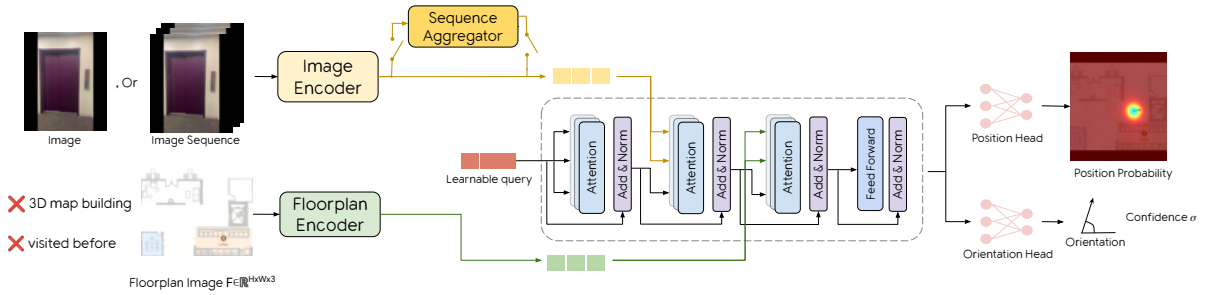

Fig. 2: **The SLoFT architecture.** A dual-encoder network based on DINO first encodes the query image and the floorplan into feature tokens. A Transformer decoder then fuses this cross-modal information. Finally, prediction heads use the resulting features to output a distribution over the camera's 2D location and yaw.

$t_q \in \mathbb{R}^{K_q \times C}$ aggregate cues from both modalities by attending to image tokens t_I and floorplan tokens t_F via multi-head cross-attention. The attention layers are proceeded by layer normalization and followed by dropout, with feed-forward network (FFN) at the end. These updated tokens, now containing a representation of the cross-modal correspondence, are then passed to the prediction heads. The choice of using learnable queries is justified by our ablation experiments.

D. Prediction heads

The final prediction of the localization is accomplished through two prediction heads: a position head that generates a localization probability map, and an orientation head that predicts the camera's yaw with its uncertainty. The updated 1D query tokens t_q are first converted to 2D feature maps $T \in \mathbb{R}^{C \times H'' \times W''}$. T then passes through a series of upsampling blocks, each composed of bilinear interpolation followed by 3x3 convolutional layers with Batch norms and ReLU to produce a single-channel heatmap $H \in \mathbb{R}^{1 \times H' \times W'}$ that aligns with the input floorplan's spatial dimensions. The camera's 2D position $p = (x, y)$ is then determined by the coordinate with the highest probability in H.

Meanwhile, the orientation head processes the same tokens t_q to regress the camera's yaw. The tokens are passed through MLP layers that yields two outputs: a 2D vector for the orientation and a scalar for the confidence score, with activation of Tanh and Sigmoid functions, respectively. To handle the periodic nature of angles, we predict a continuous representation $\bar{\mathbf{d}} = [\cos\theta, \sin\theta]$, instead of the angle θ itself. We normalize it such that $\mathbf{d} = ||\bar{\mathbf{d}}||_2$ to ensure it lies on the unit circle. The final yaw angle θ is recovered by $\theta = \tan^{-1}(\frac{\sin\theta}{\cos\theta})$, constraining the output to the range [-π, π] We use the scalar output as a confidence score σ, which is interpreted as the concentration parameter of a von Mises distribution [30], as it is the Gaussian distribution for periodic variables.

E. Sequence model

While our model is designed for effective localization from a single image, certain locations that contains few semantic cues can be ambiguous. To resolve such cases, SLoFT can optionally leverage the context provided by an image sequence. Unlike prior works that require known relative poses [7], [11], our model is capable of processing raw

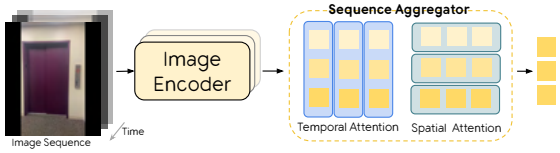

Fig. 3: **Optional Sequence Aggregator.** It alternates spatial and temporal self-attentions. The final frame's token, enriched with context from the preceding sequence, is passed to the fusion module.

image sequences directly. Given an image sequence $(I_i)_{i=1}^{M}$, each frame is first processed independently by shared image encoder, yielding M sets of feature tokens $(t_{I_i})_{i=1}^{M} \in \mathbb{R}^{M \times K \times C}$. We aggregate the spatio-temporal information by alternating between spatial attention across K dimension and temporal attention across M dimension. After that, we obtain the feature tokens from the final timestep, which now contain information accumulated from the image sequence. These contextually-enhanced tokens are then passed to the fusion module to be integrated with the floorplan features.

IV. Training

A. Training losses

SLoFT is trained end-to-end in a supervised manner using the following loss:

$$\mathcal{L} = \mathcal{L}_H + \lambda_{\mathbf{d}}\mathcal{L}_{\mathbf{d}} + \lambda_{H_{\text{smooth}}}\mathcal{L}_{H_{\text{smooth}}} + \lambda_{\hat{\mathbf{d}}_{\text{norm}}}\mathcal{L}_{\hat{\mathbf{d}}_{\text{norm}}} \quad (2)$$

, where λ balances between loss terms. The heatmap loss $\mathcal{L}_H$ supervises the probability heatmap of camera position. The ground-truth heatmap $\hat{H}$ is generated by applying a 2D Gaussian kernel centered at the ground-truth position $\hat{\mathbf{p}}$, given by $\hat{H}(\mathbf{x}) = \exp(-\frac{||\mathbf{x}-\hat{\mathbf{p}}||^2}{2\tau^2})$, where $\mathbf{x}$ is pixel coordinates in the heatmap and τ is a constant of Gaussian standard deviation. We found that the ground-truth probability heatmap should contains a small area of high probability, while the vast majority of the heatmap consists of small probability. This imbalance motivates us to define $\mathcal{L}_H$ using the Focal Loss [31] and adapt it for our soft ground-truth heatmap: $\mathcal{L}_H = \alpha(\hat{H})(1 - e^{-\text{BCE}(H,\hat{H})})^{\gamma} \cdot \text{BCE}(H, \hat{H})$, where BCE is binary cross entropy and $\alpha(x) = x \cdot \alpha \cdot (1-x)(1-\alpha)$. α and γ are constants.

For the orientation loss $\mathscr{L}_{\mathbf{d}}$, we adopt a probabilistic approach, modeling the orientation as a von Mises [30] distribution, which is a circular analogy of the normal distribution for periodic data, to capture prediction uncertainty of orientation. SLoFT predicts the distribution's mean angle $\mu = \tan^{-1}(\frac{\sin\theta}{\cos\theta})$ and concentration parameter σ. The loss is then the negative log-likelihood of the ground-truth angle $\hat{\theta}$: $\mathscr{L}_{\mathbf{d}} = -\log \mathscr{P}_{\mathscr{VM}}(\hat{\theta}|\mu,\sigma)$, where $\mathscr{P}_{\mathscr{VM}}$ is the von Mises probability density function.

To prevent noisy, isolated peaks in the position heatmap, which can harm the final position because of the argmax prediction, we add a smoothness constraint $\mathscr{L}_{H_{\text{smooth}}}$. This loss penalizes the discrete Laplacian of the heatmap, encouraging a smoother distribution: $\mathscr{L}_{H_{\text{smooth}}} = ||\nabla^2 H||_1$. To further stabilize the training of the orientation head, we encourage the orientation vector before normalization to be on the unit circle either, as: $\mathscr{L}_{\bar{\mathbf{d}}_{\text{norm}}} = (||\bar{\mathbf{d}}|| - 1)^2$ The benefits of these regularization terms for the training curve are validated in our experiments section.

B. Training data

To train our model for 3DoF localization across a wide range of scenarios, we curated a large and diverse dataset comprising 1,239,768 images and 744,304 floorplans. This collection integrates several existing datasets - including MGL [11], [32], Structured3D [33], and self-collected SLoFT dataset. This covers outdoor of walking and driving, indoor, and both real and synthetic environments. We follow the official train-validation-test splits of each public dataset and performed a standardization process: all floorplans were converted into a RGB image format, and their corresponding ground-truth poses were aligned to the pixel space of floorplans. In addition, we utilize their official pixel-per-meter (ppm), to convert the pixel scale into the metric scale.

However, floorplans in existing public datasets are often over-simplified, consisting mainly of lines and few predefined classes of areas and nodes [11], [33], which creates a significant domain gap with real-world floorplans that are rich in symbols and text. Furthermore, existing indoor datasets are typically either synthetic [5], [33]–[35] or limited to residential spaces [33], [35], while others focus exclusively on outdoor environments [32], [36]. Therefore, to address the limitations, we collected SLoFT dataset to capture the complexities exhibited in the floorplan found in real-world.

The dataset was captured across 64 floors in different office buildings. Each floor contains multiple sub-areas (e.g., lounges, rooms) and was recorded under diverse lighting conditions through windows (day and night) with transient objects. In total, it comprises 424,152 video frames of in total 4 hours. We enforce a strict data split where the building floors for the training, validation, and test sets are mutually exclusive. To generate the 3DoF poses used for training, a high-precision SLAM pipeline is adopted to provide the image pose within the floorplan. All annotations underwent a manual filtering process to remove inaccuracies caused by tracking failure. The floorplan size in metric scale is ranging from $56.7m^2$ to $344.5m^2$, with an ppm of $33.5\text{px}/m$. This variable scale of the captured area is crucial; as all floorplan images are resized to 518 during training, the model is naturally exposed to a diverse spectrum of ppm, enhancing its robustness to area size changes.

V. Experiments

A. Implementation details

To accommodate a patch size of 14, all floorplan images are resized to a uniform resolution of 518 and query image to 630. During resizing, the original aspect ratio is maintained by applying zero-padding to the shorter dimension. Ground-truths and ppm are transformed accordingly to reflect the resizing. Our sequence model processes any number of images per sequence, with a maximum of 7. The embedding dimension C is set to 768, and the number of tokens from the image, floorplan encoders, and learnable query, K_I, K_F, K_q are 45, 37, 1, respectively. Our dual encoder utilizes the DINOv2-base model with register [27], leveraging its patch tokens. The base SLoFT model contains 200M parameters, of which 40M are trainable; the optional sequence model adds another 9M parameters.

The model is trained for 83K iterations using AdamW optimizer with a weight decay of 0.01. We employ a cosine annealing scheduler, with the learning rate peaking at 1e-4 after 10K iterations and decreasing to a minimum of 1e-5. Training was conducted on 8 A100-40GB GPUs for 4 days, with a batch size of 168 and 20 CPU workers. For our loss function, the position focal loss parameters α and γ are set to 0.25 and 2.0, respectively. The loss weight for orientation, $\lambda_{\mathbf{d}}$ and orientation norm, $\lambda_{\hat{\mathbf{d}}_{\text{norm}}}$ are empirically set to 2e-4, and 1.6e-4. The weight for the position smoothness, $\lambda_{H_{\text{smooth}}}$, is initialized at 0.5 and linearly annealed to 0 during the first half of training to allow the model to produce sharper heatmaps in later stages. We apply aggressive augmentations. Query images go through color jittering, random erasing up to 30% of the image, random roll and pitch augmentation up to 20°, and random z-axis translation up to 10% of the image height. Floorplan images are augmented with random 2D rotation up to 20° and 2D translation up to 10% of floorplan size. All ground-truth labels and ppm are transformed accordingly to reflect these augmentations.

B. Evaluation metrics

The position error is the L2 pixel distance converted to meters via ppm. The heatmap representation of SLoFT can handle positions outside of the floorplan's spatial dimension; however, since the final position is determined by argmax, such out-of-bounds are excluded during evaluation. The orientation error is the absolute angular difference in degrees, and we ensured correct handling of angular periodicity. We report recall (R@, % omitted) as the percentage of predictions where errors fall within thresholds. The thresholds are chosen to each dataset's characteristics, and for Structured3D, we stick to the popular benchmark [5], [7], [10].

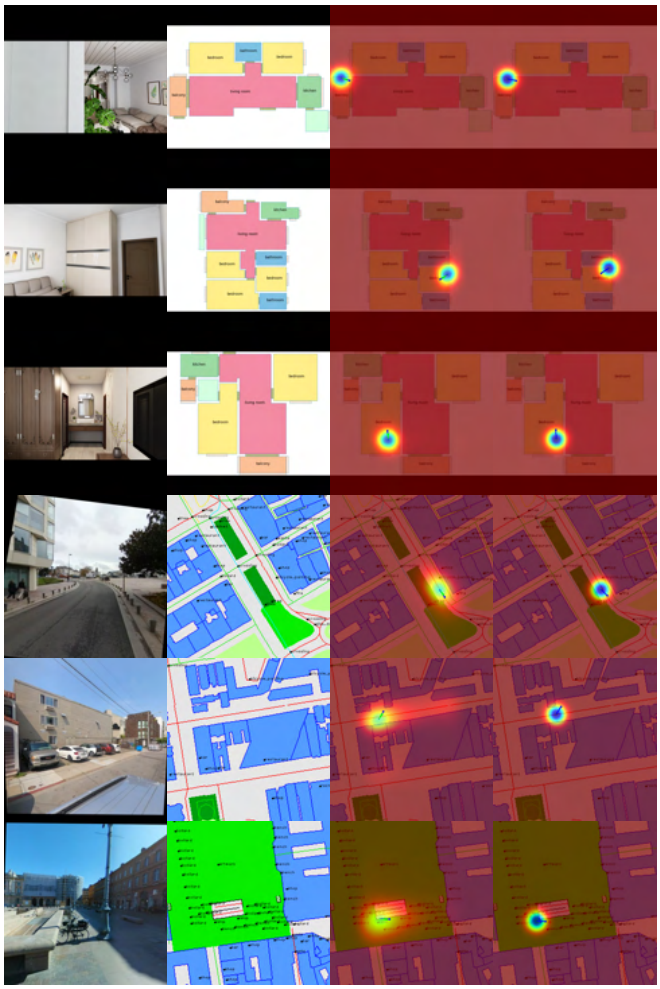

Fig. 4: **Qualitative results of SLoFT**. Our model is evaluated on the official Structured3D test set of **areas never seen during training**, and on the MGL validation set for fair comparison with previous works [11]. The results show that our model excels at understanding both the semantic context from the query image (rows 1,4,6) and implicitly reasoning about the geometry of walls and buildings (row 2,3,5), without explicit geometric supervision.

C. Training a single model

Unlike previous works that train on individual dataset or specific domains (e.g., outdoor and OSM-style maps), we train a single model, SLoFT-U, on a unified curation of four distinctive datasets spanning indoor, outdoor, public, and private areas. The balance between these datasets was determined empirically, with further details available in the Supplementary video. Remarkably, this unified model demonstrates significantly better generalization to areas unseen during training, even compared to its dataset-specific counterparts, as shown in Tab. I and Tab. IV. This underscores our model's scalability and marks a significant step forward for generalization in floorplan-based localization.

TABLE I: **Evaluation on the S3D dataset.** Our method achieves significant leap over existing works, which is further improved by leveraging text (closed-vocabulary) written on floorplan image.

	R@0.1m	R@0.5m	R@1m	R@1m30°
PF-Net [5]	0.2	1.3	3.2	0.9
LASER [10]	0.7	6.4	10.4	8.7
F3Loc [7]	1.5	14.6	22.4	21.3
SLoFT w/o text	2.4	<u>34.8</u>	<u>61.6</u>	26.8
SLoFT	<u>2.6</u>	33.8	59.7	<u>57.7</u>
SLoFT-U	**3.2**	**39.2**	**63.6**	**59.7**

D. Localization by floorplan image

We evaluate our model on the test set of Structured3D (S3D), which is a disjoint set of houses from the training data, shown at Tab. I. SLoFT removes the restrictive input requirements of prior methods (e.g., wall-only floorplans [7]–[9], point clouds [10], or upright camera poses [5]). Despite this flexibility, our model consistently outperforms the state-of-the-art. Moreover, it implicitly enables reading of text on the floorplan for a performance boost, which previous works cannot. This is also analyzed qualitatively at Fig. 5 Surprisingly, incorporating out-of-domain datasets into training (SLoFT-U) further improves S3D upon S3D-only models (all baselines and SLoFT) in 1m and 1m30° recall, signifying its scalability and generalization.

As suggested in Fig. 4, we speculate that our model implicitly leverages spatial layout cues (e.g., room shapes) to resolve ambiguities, such as distinguishing between multiple similar bedrooms (row 2,3), despite the absence of explicit geometric supervision.

E. Ablation study

We validate our key design choices in ablation studies, summarized at Tab. II. Training curves of each ablation is provided in Supplmentary. First, we assess our dual encoder, confirming our frozen DINO outperforms both fine-tuning DINO and a custom CNN. The CNN baseline consists of 11 conv layers with residual connections, batch norms, and ReLU (not at the last layer). Its final feature map, at 1/32 of the input resolution, is flatten into tokens with 2D positional encoding added. Next, we show our leranable query is more effective than standard cross-attention where image tokens serve as queries and floorplan tokens as keys and values. We hypothesize this is less suited for our sparse localization task of only one (x, y, yaw) output per input. We also analyze orientation uncertainty, by replacing our orientation loss term with $\mathscr{L}_{\mathbf{d}} = ||\mathbf{d} - \hat{\mathbf{d}}||_2^2$, where $\mathbf{d} = [\sin\theta, \cos\theta]$, which degrades both position and orientation due to the shared fusion module. Finally, ablating the regularization terms $\mathscr{L}_{H_{\text{smooth}}}$ and $\mathscr{L}_{\mathbf{d}_{\text{norm}}}$, confirms their positive impact. These results collectively justify the proposed design of SLoFT.

F. Application1: City-scale outdoor localization

Unlike prior domain-specific methods, our approach seamlessly extends from indoor to city-scale outdoor localization. To achieve this, we leverage readily available GPS data to

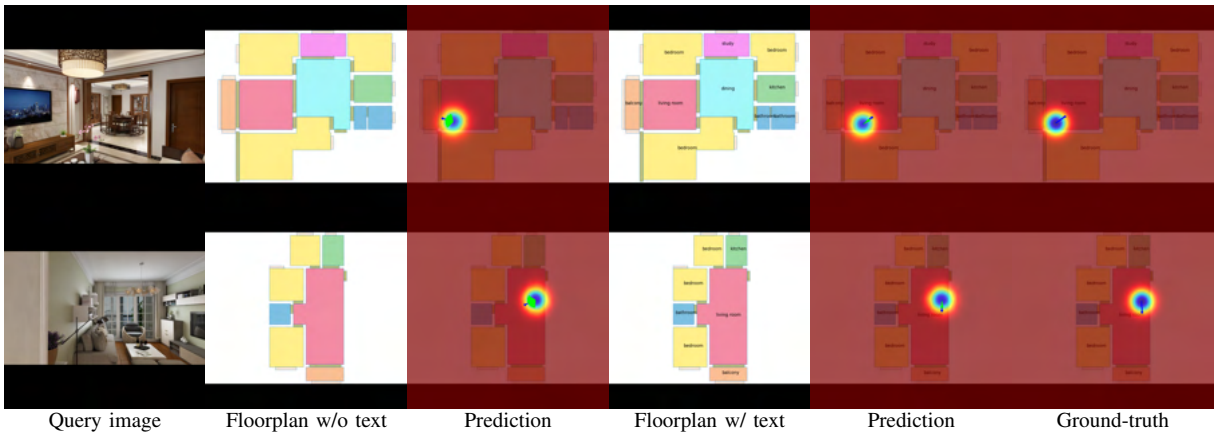

Fig. 5: **Impact of text labels on localization.** Without text, SLoFT can still approximate a plausible position from room shape. However, the visual presence of text labels in ambiguous areas (e.g., 'dining', 'balcony') provides discriminative context needed for more precise orientation.

TABLE II: **Ablation study validates our design choices.** (a) Finetune DINO, (b) CNN as encoder, (c) w/o learnable queries, (d) w/o uncertainty, (e) w/o $\mathscr{L}_{H_{\text{smooth}}}$, (f) w/o $\mathscr{L}_{\tilde{\mathbf{d}}_{\text{norm}}}$, (g) SLoFT (Full)

	(a)	(b)	(c)	(d)	(e)	(f)	(g)
R@1m30° ↑	11.5	21.7	1.4	17.1	25.1	32.2	**57.7**
xy err [m] ↓	2.37	**1.78**	5.83	2.15	1.8	1.83	1.79
yaw err [°] ↓	78.0	67.0	85.5	60.4	58.1	42.1	**19.6**

TABLE III: **Evaluation on the MGL dataset.** Our models has advantage at yaw err, while the full error distribution of ours and [11] are significantly different (analysis and justifications available at Supplementary). SLoFT-U enables both indoors and outdoors as a single model. Average xy and yaw errors are in meter and degree. R is for recall.

	xy err↓	yaw err↓	R15m↑	R30m↑	R15°↑	R30°↑	R30m30°↑
[11]	**10.42**	22.96	**77.3**	**87.7**	**83.0**	**84.7**	**79.7**
SLoFT	15.25	**19.52**	65.1	84.7	71.7	83.0	72.5
SLoFT-U	16.28	20.83	62.7	83.6	71.1	81.7	71.0

crop a local area, adding random noise to the coordinates to prevent the model from learning a center bias. For a fair comparison with the baseline, we evaluate on the validation set of the MGL dataset. As in Tab. III, our single unified model – the same model is used for indoor, outdoor, synthetic, and OSM-style floorplans – achieves performance comparable to the SOTA trained exclusively on outdoor data.

G. Application2: Real-world localization from hand-mounted phone

We extend our work to real-world experiments using the strict test set of SLoFT dataset, which features unseen building floors with challenges like dynamic objects, variable lighting, and complex floorplans. To further test our model's robustness, we isolated two particularly difficult subsets: an "ambiguous set" (e.g., symmetric areas, repeated similar areas) and a "few-shot set" with scenarios rare in the training data (e.g., dark areas, inside stair room). As shown

TABLE IV: **Evaluation on SLoFT dataset.** Our generalized model, SLoFT-U significantly outperforms its dataset-specific counterpart, suggesting our model's scalability.

		R@1m	R@3m	R@5m	R@10°	R@30°	R@50°
SLoFT	Regular	1.0	30.7	66.4	11.1	27.2	36.9
SLoFT-U	Regular	3.0	48.9	67.5	20.3	28.2	47.6
	Ambiguous	0.8	13.4	35.8	15.1	25.8	33.4
	Few-shot	5.8	11.0	37.4	44.4	58.2	68.9

in Tab. IV, our single model SLoFT-U achieves a more substantial performance gain on this challenging real-world data, compared to SLoFT. The model performs reasonably in the ambiguous and few-shot scenarios, which can be further improved by our Sequence model at Tab. V. Unfortunately, we are unable to show visual examples at this moment, due to confidential agreement.

H. Sequence aggregator

Our model, while designed for single images, can also leverage sequential data to resolve inherent ambiguities. Crucially, unlike prior methods [7], [11], our approach does not require known camera poses. We introduce SLoFT-seq, an end-to-end trained extension featuring a dedicated Sequence Aggregator module. This design allows the model to harness the rich semantic features of our base architecture, enabling it to interpret image sequences even without visual overlaps, as shown in Fig. 6. Trained on variable-length sequences, a single SLoFT-seq model adeptly handles diverse inputs at inference. Quantitative evaluations on the SLoFT dataset confirm our model's effectiveness, demonstrating improved overall performance and significant gains in ambiguous scenarios (Tab. V).

I. In-the-wild Scenarios

We test our model's in-the-wild capabilities by capturing new pairs of images and floorplans within a building in a different geographical location and different camera from the training data. As shown in Fig. 7, we observe the generalization potential of SLoFT, allowing floorplan rotations, lighting

Fig. 6: **Understanding image sequence of variable lengths.** SLoFT-seq helped positional ambiguity on a long road by recalling a previously observed exit (top), and maintains correct orientation despite misleading text by understanding the continuity of facing a building, without overlap (bottom). This was achieved without known relative camera poses.

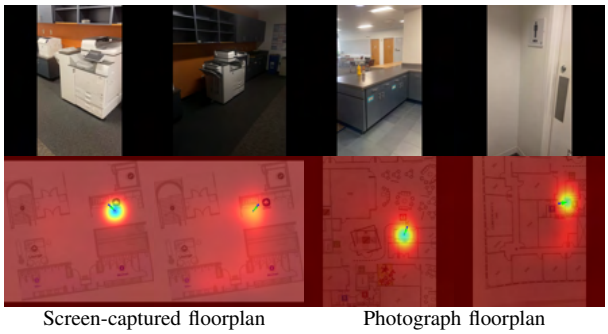

Fig. 7: **Feasibility of in-the-wild.** We conduct a preliminary study on data captured at a building unseen during training with a different camera. This setup investigates the model's capability to handle domain gaps beyond the training distribution.

TABLE V: **Effectiveness of the Sequence aggregator.** Our sequence model, SLoFT-seq, demonstrates advantage in mitigating ambiguity and overall performance.

		R@1m	R@3m	R@5m	R@10°	R@30°	R@50°
SLoFT	Ambi-guous	2.9	14.4	30.4	21.1	34.1	43.1
SLoFT-seq		**3.7**	**18.8**	**35.6**	**21.2**	**36.9**	**46.5**
SLoFT	Total	2.1	18.3	**41.2**	19.7	34.7	44.1
SLoFT-seq		**2.5**	18.3	39.0	**21.9**	**40.0**	**49.5**

changes, and different methods of capturing floorplan images (screen capture vs. photograph). The quantitative evaluation is provided at Tab. VI. We use the same SLAM pipeline for training data generation to generate ground truth. We report the xy position error as a ratio of the map image's longer side, because some in-the-wild floorplans do not provide accurate pixel-per-meter conversion. The qualitative results are shown in the supplementary video.

VI. Discussion

Interpretability We investigated the interpretability of SLoFT-U by visualizing its multi-head attention weights from the Fusion module. We refer readers to the Supplementary video for the visualizations.

TABLE VI: **Evaluation on in-the-wild** are averaged over 9 in-the-wild sequences, captured in unseen buildings with unseen cameras. xy errors are ratio of map's longer side.

xy err↓	yaw err↓	R@0.3	R@0.5	R@1.0	R@10°	R@30°	R@50°
0.216	68.99	82.2	99.8	1.0	12.9	27.6	41.4

TABLE VII: **Comparison with concurrent works on S3D dataset.** Despite being trained on a broader collection of datasets, SLoFT-U achieves performance comparable to concurrent works specialized only on S3D dataset, maintaining high competitiveness without overfitting to a single dataset.

	R@0.1m	R@0.5m	R@1m	R@1m30°
SemRayLoc [9]	5.7	**45.5**	58.8	57.5
Chen et al. [8]	**6.4**	28.6	56.9	25.2
UnLoc [37]	5.3	33.9	38.8	37.6
SLoFT	2.6	33.8	59.7	57.7
SLoFT-U	3.2	39.2	**63.6**	**59.7**

Runtime and memory During inference, our model operates at 13.3 FPS (0.075s) on a single A100 GPU, with a peak GPU memory usage of 2.04 GB. For sequential inputs, the model processes the maximum sequence length at once in 0.08s and uses 2.39 GB.

Concurrent works Tab. VII compares our model's performance to concurrent works on the S3D dataset. Our model remains competitive without being narrowly tailored to a specific dataset or floorplan topology, highlighting its potential for real-world application.

Failure cases As shown in Fig. 8, localization can fail in environments with a long street that lacks distinctive

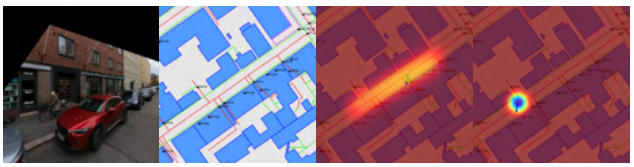

Fig. 8: **Failure cases**. Although our model may fail in extreme cases, model can quantify its confidence, allowing to flag ambiguity.

cues on the left, if the agent only sees the left. However, the probabilistic nature of our outputs allows the system to quantify its confidence and prompt the user for assistance.

Future integration with navigation The predicted confidence can be integrated with downstream tasks by allowing systems to set application-specific thresholds. For instance, pedestrian navigation may tolerate lower confidence, whereas autonomous robots can restrict to high confidence or trigger active exploration for reduced uncertainty.

VII. Conclusion

We introduce SLoFT, an end-to-end neural network that outperforms significantly from state-of-the-art floorplan-based visual localizations. Our model demonstrates versatility across indoor, outdoor, real, and synthetic environments without requiring known relative camera poses, depths, or constraints on roll and pitch. SLoFT learns to leverage crucial semantic cues from rich visual floorplans. Experiments show the model's understanding of both structures and semantics in previously unseen environments. We believe this work, by removing the barriers of 3D map creating and maintenance, represents an important step towards a scalable and simple localization system based on readily available floorplan images that mimics the human ability to intuitively ground oneself in an environment by looking at a 2D floorplan.

Acknowledgement We would like to thank Penghe (Adam) Zu, Kan Huang, Thanh Vu, Dongxu Zhao, and Ramin Nakhli for helping experimental setups and Mingxi Jia and Yichen Wei for helping Spot robot operation.

References

[1] Y. Li, *et al.*, "Location recognition using prioritized feature matching," in *European conference on computer vision*. Springer, 2010, pp. 791–804.

[2] T. Sattler, *et al.*, "Fast image-based localization using direct 2d-to-3d matching," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 667–674.

[3] G. Lowe, "Sift-the scale invariant feature transform," *Int. j*, vol. 2, no. 91-110, p. 2, 2004.

[4] R. Arandjelovic, *et al.*, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[5] P. Karkus, *et al.*, "Particle filter networks with application to visual localization," in *Conference on robot learning*. PMLR, 2018, pp. 169–178.

[6] H. Howard-Jenkins and V. A. Prisacariu, "Lalaloc++: Global floor plan comprehension for layout localisation in unvisited environments," in *European Conference on Computer Vision*. Springer, 2022, pp. 693–709.

[7] C. Chen, *et al.*, "F3loc: fusion and filtering for floorplan localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 029–18 038.

[8] B. Chen, *et al.*, "Perspective from a broader context: Can room style knowledge help visual floorplan localization?" *arXiv preprint arXiv:2508.01216*, 2025.

[9] Y. Grader and H. Averbuch-Elor, "Supercharging floorplan localization with semantic rays," *arXiv preprint arXiv:2507.09291*, 2025.

[10] Z. Min, *et al.*, "Laser: Latent space rendering for 2d visual localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 122–11 131.

[11] P.-E. Sarlin, *et al.*, "Orienternet: Visual localization in 2d public maps with neural matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 632–21 642.

[12] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[13] S. Lynen, *et al.*, "Get out of my lab: Large-scale, real-time visual-inertial localization." in *Robotics: Science and Systems*, vol. 1, no. s 1, 2015.

[14] L. Zhang, *et al.*, "Visual localization in 3d maps: Comparing point cloud, mesh, and nerf representations," *arXiv preprint arXiv:2408.11966*, 2024.

[15] J. Miao, *et al.*, "A survey on monocular re-localization: From the perspective of scene map representation," *IEEE Transactions on Intelligent Vehicles*, 2024.

[16] L. Schmid, *et al.*, "Khronos: A unified approach for spatio-temporal metric-semantic slam in dynamic environments," in *Proc. of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[17] N. Hughes, *et al.*, "Hydra: A real-time spatial perception system for 3D scene graph construction and optimization," in *Robotics: Science and Systems (RSS)*, 2022.

[18] J. C. Virgolino Soares, *et al.*, "Visual localization and mapping in dynamic and changing environments," *Journal of Intelligent & Robotic Systems*, vol. 109, no. 4, p. 95, 2023.

[19] L. Schmid, *et al.*, "Dynablox: Real-time detection of diverse dynamic objects in complex environments," vol. 8, no. 10, pp. 6259–6266, 2023.

[20] D. Rosen, *et al.*, "Towards lifelong feature-based mapping in semi-static environments," in *IEEE International Conference on Robotics and Automation*, Stockholm, Sweden, May 2016, pp. 1063–1070.

[21] F. Nobre, *et al.*, "Online probabilistic change detection in feature-based maps," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 3661–3668.

[22] A. Kendall, *et al.*, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.

[23] S. Zhu, *et al.*, "Vigor: Cross-view image geo-localization beyond one-to-one retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3640–3649.

[24] H. Howard-Jenkins, *et al.*, "Lalaloc: Latent layout localisation in dynamic, unvisited environments," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 107–10 116.

[25] N. Samano, *et al.*, "You are here: Geolocation by embedding maps and images," in *European Conference on Computer Vision*. Springer, 2020, pp. 502–518.

[26] J. Li, *et al.*, "Flona: Floor plan guided embodied visual navigation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 14, 2025, pp. 14 610–14 618.

[27] M. Oquab, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[28] A. Dosovitskiy, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[29] N. Carion, *et al.*, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[30] H. Risken, "Fokker-planck equation," in *The Fokker-Planck equation: methods of solution and applications*. Springer, 1989, pp. 63–95.

[31] T.-Y. Lin, *et al.*, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[32] G. Neuhold, *et al.*, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4990–4999.

[33] J. Zheng, *et al.*, "Structured3d: A large photo-realistic dataset for structured 3d modeling," in *European Conference on Computer Vision*. Springer, 2020, pp. 519–535.

[34] B. Shen, *et al.*, "igibson 1.0: A simulation environment for interactive tasks in large realistic scenes," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 7520–7527.

[35] S. Cruz, *et al.*, "Zillow indoor dataset: Annotated floor plans with 360deg panoramas and 3d room layouts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2133–2143.

[36] A. Geiger, *et al.*, "Vision meets robotics: The kitti dataset," *The international journal of robotics research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[37] M. Wüest, *et al.*, "Unloc: Leveraging depth uncertainties for floorplan localization," *arXiv preprint arXiv:2509.11301*, 2025.