

# Debiased Multiplex Tokenizer for Efficient Map-Free Visual Relocalization

Wenshuai Wang<sup>1,2</sup>, Hong Liu<sup>1,\*</sup>, Shengquan Li<sup>2</sup>, Peifeng Jiang<sup>1</sup>, Runwei Ding<sup>2,\*</sup>

<sup>1</sup>State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School

<sup>2</sup>Pengcheng Laboratory, Shenzhen, China

wangws@stu.pku.edu.cn, hongliu@pku.edu.cn, dingrw@pcl.ac.cn

## Abstract

Image-based feature representation plays a critical role in visual localization, enabling robots to estimate their positions and orientations in GPS-denied environments. However, this task is often undermined by significant variations in camera viewpoints and scene appearances. Recently, map-free visual relocalization (MFVR) has emerged as a promising paradigm due to its compatibility with lightweight deployment and privacy isolation on mobile devices. In this paper, we propose the Debiased Multiplex Tokenizer (DeMT) as a novel method for versatile and efficient MFVR. Specifically, DeMT performs relative pose regression through an integrated framework built upon a pretrained vision Mamba encoder, comprising three key modules: First, Multiplex Interactive Tokenization yields robust image tokens with non-local affinities and cross-domain descriptions; Second, Debiased Anchor Registration facilitates anchor token matching through proximity graph retrieval and causal pointer attribution; Third, Geometry-Informed Pose Regression empowers multi-layer perceptrons with a gating mechanism and spectral normalization to support both pair-wise and multi-view modes. Extensive evaluations across five public datasets demonstrate that DeMT substantially outperforms existing baselines and ablation variants in diverse indoor and outdoor environments.

**Code** — <https://github.com/wwsbot/DeMT>

## Introduction

Visual localization (VL), a cornerstone of intelligent perception, precisely determines the six-degree-of-freedom (6DoF) pose of an autonomous mobile robot by referencing geo-tagged imagery or preconstructed maps. Despite its growing applications in smart homes and cities, VL encounters persistent real-world challenges, such as viewpoint ambiguity, motion blur, illumination shifts, appearance variations, dynamic occlusions (*e.g.* pedestrians or vehicles), and their complex combinations.

The standard VL pipeline comprises feature extraction, matching, and pose estimation. As Figure 1 illustrates, map-based methods represented by absolute pose estimation (APE) incur high computational overhead during 2D-3D registration. Conversely, map-free visual relocaliza-

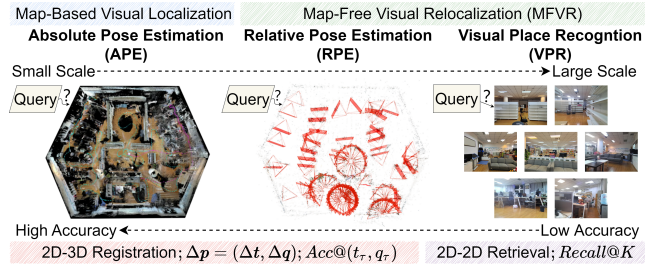


Figure 1: Main paradigms of visual localization. For a query image, absolute pose estimation (APE) matches it directly to the 3D scene map, while relative pose estimation (RPE) and visual place recognition (VPR) need to retrieve the reference dataset. Moreover, RPE is supposed to make a better trade-off between scene scale and position accuracy.

tion (MFVR) approaches including relative pose estimation (RPE) and visual place recognition (VPR) prioritize lightweight computation and scene generalization (Arnold et al. 2022). Since RGB images convey rich color, geometric, and semantic cues of a scenario, deep learning has been widely applied to visual modeling over the past decade. While convolutional neural networks (CNNs) remain limited by receptive fields (Chen et al. 2024a), vision Transformers (ViTs) leverage self-attention and cross-attention to capture global context with parallel computation (Yeh et al. 2024). Vision Mambas (ViMs) further employ state space models (SSMs) for dynamic feature filtering in linear time complexity (Zhang et al. 2024). Nonetheless, these advanced encoders can introduce inherent biases into feature representations (Wang et al. 2018).

To mitigate this issue, an effective tokenizer is essential for distilling invariant scene features into robust image tokens. This strategy unifies image retrieval and pose estimation within a relative pose regression (RPR) pipeline. Typical feature extraction techniques broadly fall into two categories: global descriptors encode entire images into comparable vectors but often neglect salient structural details, while local detectors extract fragmented patterns from distinctive pixels or patches at the cost of enumeration precision. Therefore, more discriminative descriptors must fuse their complementary strengths explicitly or implicitly.

\*Corresponding authors.

There also exists a theoretical dilemma on token matching: regression frameworks struggle to synthesize novel viewpoints (Toft et al. 2020), while a domain gap persists between VPR’s discrete outputs and RPR’s continuous embeddings despite their shared reliance on image retrieval. This discrepancy manifests as an out-of-distribution generalization problem between cameras, where causal inference offers a pathway to remove confounders in tokenization. In addition, standard multi-layer perceptrons (MLPs) lack adaptive weight modulation and geometric smoothness constraints despite their prevalence in pose regression.

In this paper, we introduce the Debiased Multiplex Tokenizer (DeMT), an efficient MFVR framework for robust sensing and fast computing. As depicted in Figure 2, DeMT integrates three core modules: (1) Multiplex Interactive Tokenization (MIT) enhances sparse voxel affinities in ViM-derived feature maps via refined non-local self-attention (Zeng et al. 2024), concurrently addressing spatial sparsity, frequency consistency, and channel redundancy for robust feature aggregation; (2) Debiased Anchor Registration (DAR) employs hierarchical navigable small world (HNSW) graphs (Malkov and Yashunin 2018) for incremental retrieval of proximate tokens, and then creates a specific causal Mamba pointer (CMP) for their counterfactual attribution, ensuring the structural consistency with visual backbone; (3) Geometry-Informed Pose Regression (GIPR) composes a novel gated linear variant (Shazeer 2020), *i.e.* spectral swish gated MLPs (SSG-MLPs), to regress 6DoF poses from single or multiple frames, demonstrating superior convergence and generalization for supervised learning.

Fundamentally, DeMT establishes a self-attentive mapping from image homography to feature proximity and employs causal reasoning to model pose correlation. In other words, it not only incorporates frequency calibration into feature description to overcome potential interferences from motion blur or illumination changes, but also achieves linear time complexity through the integration of Mamba codec and HNSW search. Compared with a series of mainstream methods, DeMT shows superiority in feature fidelity, localization accuracy, inference speed, and scene adaptability on diverse VL datasets (Shotton et al. 2013; Taira et al. 2018; Kendall et al. 2015; Toft et al. 2020). Besides, all angular values are unified into unit quaternions during pose prediction to prevent the gimbal lock (Liu and Zhang 2023).

Our primary contributions are summarized as follows:

- We present the Debiased Multiplex Tokenizer (DeMT) for high-fidelity tokenization of scene images, leveraging multiplex interactive learning on preliminary feature maps along voxel, spatial, frequency, and channel dimensions via an enhanced self-attention mechanism.
- Both tailored for token-pose debiasing, CMP efficiently scores multi-view correspondences for anchor emergence with uniform linear time complexity, while SSG-MLPs ensure precise and practical pose regression aligned with  $SE(3)$  robot kinematics.
- Comprehensive experiments on various scenes confirm DeMT’s efficacy in both indoor and outdoor environments, even under challenging circumstances.

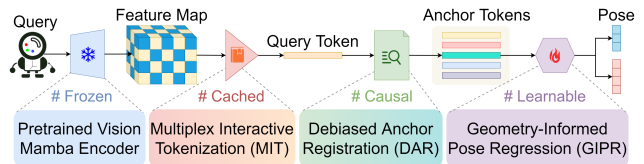


Figure 2: The pipeline of our proposed method. Four modules are cascaded for relative pose regression (RPR).

## Related Work

### Visual Place Recognition

VPR is typically formulated as an image retrieval problem based on similarity thresholds. The pioneering method NetVLAD (Arandjelovic et al. 2016) integrates CNNs with classical VLAD descriptors (Jégou et al. 2010) for global feature aggregation. Subsequent advancements evolve along two main paths. Explicit methods, such as DOLG (Yang et al. 2021), Patch-NetVLAD (Hausler et al. 2021), and TransVPR (Wang et al. 2022), enhance robustness through local feature concatenation. In contrast, CosPlace (Berton et al. 2022) and MixVPR (Ali-Bey et al. 2023) explore implicit fusion techniques to mitigate feature redundancy. Recent efforts also adapt pre-trained DINOv2 (Oquab et al. 2023) for self-supervised feature representation.

### Absolute Pose Estimation

APE aims to establish precise 2D-3D correspondences through map structures or neural encodings. Structure-based pipelines (SbP) (Sattler et al. 2017; Giang et al. 2024) iteratively resolve 2D-3D matches, albeit with high computational overhead. Deep learning has led to two main streams: scene coordinate regression (SCR) and absolute pose regression (APR), exemplified by the DSAC (Brachmann and Rother 2021) and PoseNet (Kendall et al. 2015) families, respectively. However, their effectiveness in dynamic environments is limited by the high cost of map construction and the sensitivity to training data coverage. SACNet (Wang et al. 2024a), LENS (Moreau et al. 2022), and Maprepo (Chen et al. 2024b) address these challenges by focusing on feature compression, synthetic data generation, and scene generalization, respectively.

### Relative Pose Estimation

RPE recovers the 6DoF pose of a query image from similar reference images, offering greater flexibility and generalizability. Unlike conventional hierarchical models (Sarlín et al. 2019) that heuristically solve the essential matrix, Relative Pose Regression (RPR) directly predicts pose errors between image pairs in a data-driven manner. AnchorNet (Saha et al. 2018) extends PoseNet by clustering anchor frames to improve robustness. NN-Net (Laskar et al. 2017) learns pairwise differences using Siamese networks, while ReLocNet (Balntas et al. 2018) and EssNet (Zhou et al. 2020b) improve upon this paradigm through continual learning and geometric constraints, respectively. RelFormer (Idan et al. 2024) employs a dual-branch Transformer to address scene generalization challenges. RelPoseGNN (Turkoglu

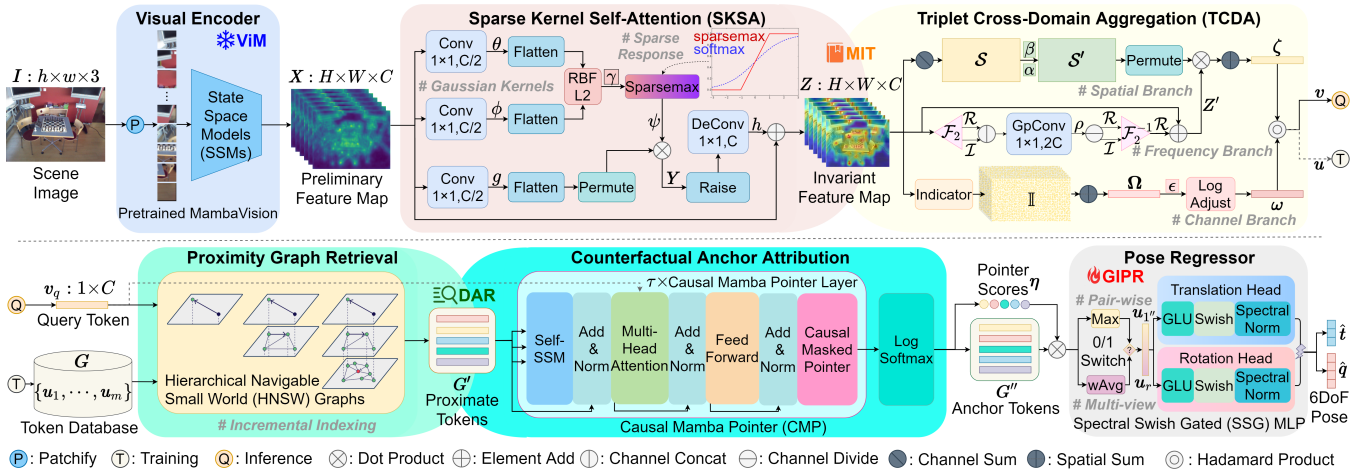


Figure 3: The overview of Debiased Multiplex Tokenizer (DeMT). This framework sequentially performs self-attentive mapping from images to tokens and causal inference to poses. RBF, DeConv, and GpConv stand for radial basis function, deconvolution, and grouped convolution, respectively.

et al. 2021) exploits multi-frame collaboration using graph neural networks (GNNs). ReLoc3r (Dong et al. 2025) incorporates multi-view epipolar geometry constraints, achieving state-of-the-art (SoTA) indoor RPR performance but generalizes poorly to outdoor environments.

## Methodology

### Feature Map Encoding

Given a scene image  $I \in \mathbb{R}^{h \times w \times 3}$ , we adopt a pretrained ViM model  $f_{\text{vim}}(\cdot)$  to extract preliminary feature map  $X \in \mathbb{R}^{H \times W \times C}$  as  $X = f_{\text{vim}}(I)$ , where  $H$ ,  $W$ , and  $C$  denote height, width, and channel dimensions, respectively.

### Multiplex Interactive Tokenization

MIT aggregates  $X$  into a unique token  $v \in \mathbb{R}^C$  through two core complements: Sparse Kernel Self-Attention (SKSA) and Triplet Cross-Domain Aggregation (TCDA).

**Sparse Kernel Self-Attention.** SKSA extends vanilla self-attention via Gaussian kernels and sparse responses, incorporating a voxel residual for feature enhancement.

**Gaussian Kernels.** To capture invariant non-local dependencies, we replace the dot-product attention with Gaussian kernels based on exponential Euclidean distance, which is computed with affinity kernels  $\theta(\cdot)$  and  $\phi(\cdot)$  as:

$$f_{\text{gk}}(\mathbf{X}) = \exp(-\gamma \|\mathbf{W}_\theta \mathbf{X} - \mathbf{W}_\phi \mathbf{X}\|^2), \quad (1)$$

where  $\mathbf{W}_\theta$  and  $\mathbf{W}_\phi$  are their learnable projection matrices, and  $\gamma > 0$  is a scaling factor.

**Sparse Response.** To obtain a more sparse affinity matrix, sparsemax (Martins and Astudillo 2016) activation is used to prevent overfitting compared to softmax, yielding the SKSA response  $\mathbf{Y} \in \mathbb{R}^{H \times W \times C}$ :

$$\mathbf{Y} = \max\{0, f_{\text{gk}}(\mathbf{X}) - \psi[f_{\text{gk}}(\mathbf{X})\mathbf{J}_{C,1}]\}g(\mathbf{X}), \quad (2)$$

where  $\psi(\cdot)$  denotes a cutoff threshold, and  $g(\cdot)$  is a positional embedding function.  $\mathbf{J}_{i,j} = \mathbf{1}_{i \times j}$  hereinafter is an all-one matrix.

At this end,  $\mathbf{Y}$  is further raised to  $\mathbf{X}$ 's original size via inverse function  $h(\cdot) = g^{-1}(\cdot)$ , and the residual is added to obtain an invariant feature map  $\mathbf{Z} \in \mathbb{R}^{H \times W \times C}$ :

$$\mathbf{Z} = \mathbf{X} + \mathbf{Y}\mathbf{W}_h. \quad (3)$$

**Triplet Cross-Domain Aggregation.** TCDA processes  $\mathbf{Z}$  through spatial, frequency, and channel branches to generate the final token  $v$ .

**Spatial Branch.** The spatial response  $\mathbf{S} \in \mathbb{R}^{H \times W}$  is derived by channel summation as  $\mathbf{S} = \mathbf{Z}\mathbf{J}_{C,1}$ , and scaled to canonical form  $\mathbf{S}' \in \mathbb{R}^{H \times W}$  as:

$$\mathbf{S}' = \left[ \frac{\mathbf{S}}{(\mathbf{J}_{W,H}\mathbf{S}\alpha)^\frac{1}{\alpha}} \right]^\frac{1}{\beta}, \quad (4)$$

where  $\alpha$  and  $\beta$  are scaling factors.

**Frequency Branch.** Considering motion blur and light shifts, the potential noise is corrected by an adaptive frequency convolution:

$$\mathbf{Z}' = \mathcal{R}(\mathcal{F}_2^{-1}(\mathbf{W}_\rho * \mathcal{R}(\mathcal{F}_2(\mathbf{Z})), \mathcal{I}(\mathcal{F}_2(\mathbf{Z})))) + \mathbf{Z}, \quad (5)$$

where  $\mathcal{F}_2(\cdot)$  denotes 2D fast Fourier transform (Chi, Jiang, and Mu 2020) and  $\mathcal{F}_2^{-1}(\cdot)$  denotes its inverse;  $\mathcal{R}(\cdot)$  extracts real components and  $\mathcal{I}(\cdot)$  extracts imaginary ones;  $\mathbf{W}_\rho$  is the weight matrix of  $1 \times 1$  grouped convolution.

Interim weighted pooling is then performed to yield the spatial-frequency token  $\zeta \in \mathbb{R}^C$  as:

$$\zeta = \mathbf{J}_{W,H}\mathbf{S}'^\top \mathbf{Z}'. \quad (6)$$

**Channel Branch.** The channel response  $\Omega \in \mathbb{R}^C$  quantifies positive spatial activations:

$$\Omega = \frac{1}{HW} \mathbf{J}_{W,H} \mathbb{I}(\mathbf{Z} > 0), \quad (7)$$

where  $\mathbb{I}(\cdot)$  is an indicator function.

For channel normalization,  $\Omega$  is logarithmically adjusted to  $\omega \in \mathbb{R}^C$  with a stability constant  $\epsilon$ :

$$\omega = \log\left(\frac{C\epsilon + \Omega\mathbf{J}_{C,1}}{\epsilon + \Omega}\right). \quad (8)$$

Thus, the final token  $v$  of  $I$  is obtained by  $v = \zeta \odot \omega$ , where  $\odot$  means Hadamard product herein. Specifically, reference tokens are referred to as  $u$  for clarity.

### Debiased Anchor Registration

DAR estimates pose differences between pair-wise images while addressing potential misalignments through Proximity Graph Retrieval and Counterfactual Anchor Attribution.

**Proximity Graph Retrieval.** Given a query token  $v_q$  and a reference token database represented by proximity graphs  $G = \{u_1, \dots, u_m\}$ , the HNSW algorithm  $f_{\text{hnsw}}(\cdot, \cdot)$  incrementally retrieves the top- $k$  proximate candidates:

$$G' = f_{\text{hnsw}}(G, v_q) = (u_{1'}, \dots, u_{k'}), \quad (9)$$

where  $i'$  denotes the index of  $i$ -th proximate token.

**Counterfactual Anchor Attribution.** To correct the correspondences between tokens and poses, a  $\tau$ -layer CMP  $f_{\text{cmp}}^\tau(\cdot, \cdot)$  refines  $G'$  into a debiased anchor set  $G''$ :

$$G'' = f_{\text{cmp}}^\tau(G', v_q) = (u_{1''}, \dots, u_{k''}), \quad (10)$$

where  $i''$  denotes the index of  $i$ -th anchor token.

### Geometry-Informed Pose Regression

GIPR employs decoupled 2-layer SSG-MLPs to regress the relative pose difference  $\Delta p = [\Delta t \in \mathbb{R}^3, \Delta q \in \mathbb{R}^4]$  as:

$$\Delta p = [f_{sn}(f_{sw}(t_q) \odot \sigma(t_q)), f_{sn}(f_{sw}(q_q) \odot \sigma(q_q))], \quad (11)$$

where  $\sigma(\cdot)$  denotes sigmoid activation,  $f_{sw}(\cdot)$  denotes a linear layer with swish activation, and  $f_{sn}(\cdot)$  denotes another linear layer with spectral normalization (Bjorck, Gomes, and Weinberger 2021).

Therefore, for each anchor token  $v_r$  with its known pose  $p_r = [t_r, q_r]$ , the predicted absolute pose  $\hat{p}_q^{(r)} = [\hat{t}_q^{(r)}, \hat{q}_q^{(r)}]$  of  $I_q$  is computed by applying the regressed  $\Delta p$  to  $p_r$  as:

$$\hat{t}_q^{(r)} = t_r + \Delta t, \quad \hat{q}_q^{(r)} = q_r \otimes \Delta q, \quad (12)$$

where  $\otimes$  denotes quaternion multiplication. To avoid ambiguities, the resulting quaternion is normalized to unit length.

### Loss and Inference

**Training Loss.** The RPR loss  $L_{\text{rpr}}$  is defined as pose differences by using the  $L_1$  norm:

$$L_{\text{rpr}} = \|\hat{t}_q - t_q\|_1 + \|2 \arccos |\langle q_q, \hat{q}_q \rangle| \|_1. \quad (13)$$

where  $\langle \cdot, \cdot \rangle$  denotes the dot product between quaternions, and  $\arccos(\cdot)$  returns the angle in radians.

Overall, the total loss  $L_{\text{total}}$  is the sum of three parts:

$$L_{\text{total}} = L_{\text{rpr}} + L_{\text{hnsw}} + L_{\text{cmp}}, \quad (14)$$

where  $L_{\text{hnsw}}$  and  $L_{\text{cmp}}$  are negative log-likelihood (NLL) losses that encourage  $G'$  and  $G''$  to contain the true matching reference tokens for  $v_q$ .

**Pair-wise Inference.** Only the top-ranked anchor  $u_{1''}$  is selected from  $G''$ , so the absolute pose error is computed as:

$$E_{pw}(v_q, u_{1''}) = (\|t - \hat{t}\|_2, 2 \arccos |\langle q, \hat{q} \rangle| \frac{180^\circ}{\pi}). \quad (15)$$

**Multi-view Inference.** A structural smoothing error can be obtained by directly weighting average on multiple frames with their confidence scores  $\eta = \{\eta_1, \dots, \eta_k\} \in [0, 1]^k$  from CMP as  $E_{mv}^k(v_q, G'') = \sum_{i=1}^k \eta_i E_{pw}(v_q, u_{i''})$ .

Dataset	# ref.	# query	Motion	Light	Season	Noise
7Scenes (Shotton et al. 2013)	26.0k	17.0k	★★☆	★☆☆	☆☆☆	☆☆☆
Cambridge (Kendall et al. 2015)	8.4k	4.8k	★★★	★★☆	☆☆☆	★★★
InLoc (Taira et al. 2018)	10.0k	329	★★☆	★★★	☆☆☆	★★★
Aachenv1.1 (Toft et al. 2020)	6.7k	1.0k	★★☆	★★★	☆☆☆	★★★
Extend CMU-Seasons (Toft et al. 2020)	60.9k	56.6k	★★★	★★★	★★★	★★★

Table 1: Details of the adopted MFVR datasets.

## Experiments

### Datasets and Evaluation Metrics

**Datasets.** The proposed method is evaluated on five standard VL datasets: 7Scenes (Shotton et al. 2013), Cambridge Landmarks (Kendall et al. 2015), InLoc (Taira et al. 2018), Aachen Day-Night v1.1 (Toft et al. 2020), and Extended CMU-Seasons (Toft et al. 2020). As detailed in Table 1, they encompass both indoor and outdoor environments, and vary significantly in scale and conditions. All datasets are publicly available to ensure experimental reproducibility.

**Evaluation Metrics.** VL performance is assessed by the absolute pose errors  $E_{pw}$  (pair-wise) or  $E_{mv}$  (multi-view). For a comprehensive comparison, we report the median translation and rotation errors across test samples. Where applicable, their averages, standard deviations, and accuracy percentages under specific thresholds are also provided.

### Implementation Details

**Models.** The pretrained MbambaVision-L (Hatamizadeh and Kautz 2025) is chosen as the visual encoder with fixed feature channel dimension  $C=1536$ . Its final average pooling and fully connected layers are removed from the backbone to extract preliminary feature maps.

**Hyperparameters.** For MIT,  $\theta(\cdot)$ ,  $\phi(\cdot)$  in Equation (1) are implemented by  $1 \times 1$  radial basis function (RBF) convolution (Amirian and Schwenker 2020), while  $g(\cdot)$  in Equation (2) is by  $1 \times 1$  bottleneck convolution (Zhou et al. 2020a).  $\alpha$  and  $\beta$  in Equation (4) are set to 0.5 and 2, respectively.  $\epsilon$  in Equation (8) is set to 0.0001. For DAR, HNSW’s  $k$  is set to 10, and CMP’s  $\tau$  is 6. For GIPR, SSG-MLP’s hidden channel is set to 1024. The framework is optimized using adaptive moment estimation (Adam) with initial learning rate of 0.01, momentum of 0.9, weight decay of 0.0001, and mini-batch size of 16. Training is conducted for 30 epochs on indoor datasets and 200 epochs on outdoor datasets.

**Platform.** All experiments are performed on an Ubuntu 20.04 system equipped with an NVIDIA GeForce RTX 3090 GPU and an Intel(R) Core i5-13400F CPU, running PyTorch 1.11.7 with Python 3.8 and CUDA 12.1.

### Quantitative Comparisons

**RPR Performance on Benchmark Datasets.** Table 2 compares the RPR performance of DeMT against representative VL baselines on the 7Scenes and Cambridge Landmarks datasets. For pairwise calibration, DeMT achieves SoTA results among pair-wise RPR methods, with average errors of (7.3cm, 3.25°) for indoor scenes and (15.5cm, 1.02°) for outdoor scenes. These results underscore DeMT’s significant

Dataset	Metrics Method	7Scenes (indoor, normal)								Cambridge Landmarks (outdoor, normal)					
		Median Error (cm/°) ↓								Median Error (cm/°) ↓					
		Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Average	College	Hospital	Shop	Church	Average-4	Court
SP	AS (Sattler et al. 2017)	4/2.00	3/1.50	2/1.50	9/3.60	8/3.10	7/3.40	3/2.20	5.1/2.47	42/0.60	44/1.00	12/0.40	19/0.50	29.3/0.63	-/-
	PixLoc (Sarlin et al. 2021)	2/0.80	2/0.73	1/0.82	3/0.82	4/1.21	3/1.20	5/1.30	2.9/0.98	14/0.24	16/0.32	5/0.23	10/0.34	11.3/0.28	30/0.14
	DeViLoc (Giang et al. 2024)	2/0.78	2/0.74	1/0.65	3/0.82	4/1.02	3/1.19	4/1.12	2.7/0.90	12/0.21	13/0.28	4/0.18	7/0.23	9.0/0.23	18/0.11
	FaVoR (Polizzi et al. 2025)	1/0.20	1/0.40	1/0.60	2/0.40	1/0.30	1/0.30	6/1.60	1.9/0.50	18/0.30	27/0.50	5/0.30	11/0.40	15.3/0.38	29/0.20
SCR	DSAC* (Brachmann et al. 2021)	2/1.10	2/1.24	1/1.82	3/1.15	4/1.34	4/1.68	3/1.16	2.7/1.36	18/0.30	21/0.40	5/0.30	15/0.60	14.8/0.40	49/0.30
	ACE (Brachmann et al. 2023)	2/1.10	2/1.80	2/1.10	3/1.40	3/1.30	3/1.30	3/1.20	2.7/1.31	28/0.40	31/0.60	5/0.30	18/0.60	20.5/0.48	43/0.20
	NeuMap (Tang et al. 2023)	2/0.81	3/1.11	2/1.17	3/0.98	4/1.11	4/1.33	4/1.12	3.1/1.09	19/0.14	36/0.19	25/0.06	53/0.17	33.3/0.14	10/0.06
	HSCNet++(Wang et al. 2024b)	2/0.70	2/0.72	1/0.80	2/0.69	4/1.00	4/1.15	3/1.02	2.6/1.36	19/0.34	20/0.31	6/0.24	9/0.30	13.5/0.30	39/0.23
	D2S (Bui et al. 2024)	2/0.57	2/0.74	1/0.75	2/0.62	3/0.83	3/1.04	13/2.02	3.7/0.94	7/0.12	15/0.29	3/0.17	8/0.25	8.3/0.21	23/0.11
	SACNet <sup>†</sup> (Wang et al. 2024a)	2/0.53	2/0.71	1/0.59	3/0.65	2/0.65	3/0.93	2/0.40	2.1/0.64	17/0.30	18/0.30	6/0.30	12/0.30	13.3/0.31	47/0.32
APR	PoseNet (Kendall et al. 2015)	32/8.12	47/14.4	29/12.0	48/7.68	47/8.42	59/8.64	47/13.8	44.1/10.4	192/5.40	231/5.40	146/8.10	266/8.50	208.8/6.80	-/-
	PAE (Shavit and Keller 2022)	12/4.95	24/9.31	14/12.5	19/5.79	18/4.89	18/6.19	25/8.74	18.6/7.48	90/1.49	207/2.58	99/3.88	164/4.16	140.0/3.03	-/-
	DFNet (Chen et al. 2022)	5/1.88	17/6.45	6/3.63	8/2.48	10/2.78	22/5.45	16/3.29	12.0/3.71	73/2.37	200/2.98	67/2.21	137/4.03	119.3/2.90	-/-
	LENS <sup>†</sup> (Moreau et al. 2022)	3/1.30	10/3.70	7/5.80	7/1.90	8/2.20	9/2.20	14/3.60	8.3/2.96	33/0.50	44/0.90	27/1.60	53/1.60	39.3/1.15	-/-
	PMNet <sup>†</sup> (Lin et al. 2024)	4/1.70	10/4.51	7/4.23	7/1.96	14/3.33	14/3.36	16/3.62	10.3/3.24	68/1.97	103/1.31	58/2.10	133/3.73	90.5/2.27	-/-
Marepo (Chen et al. 2024b)	2/1.24	2/1.39	2/2.03	3/1.26	4/1.48	4/1.71	6/1.67	3.3/1.54	-/-	-/-	-/-	-/-	-/-	-/-	
RPR (pw)	NN-Net (Laskar et al. 2017)	13/6.50	26/12.7	14/12.3	21/7.40	24/6.40	24/8.00	27/11.8	21.3/9.30	-/-	-/-	-/-	-/-	-/-	-/-
	ReLocNet (Balntas et al. 2018)	12/4.10	26/10.4	14/10.5	18/5.30	26/4.20	23/5.10	28/7.50	21.0/6.73	-/-	-/-	-/-	-/-	-/-	-/-
	AnchorNet (Saha et al. 2018)	8/4.12	16/11.1	9/11.2	11/5.38	14/3.55	13/5.29	21/11.9	13.1/7.51	79/0.95	211/3.05	77/3.25	122/3.02	122.3/2.57	589/3.53
	NC-EssNet (Zhou et al. 2020b)	12/5.60	26/9.60	14/10.7	20/6.70	22/5.70	22/6.30	31/7.90	21.0/7.50	61/1.60	95/2.70	71/3.40	112/3.60	84.8/2.80	-/-
	Map-free (Arnold et al. 2022)	9/2.66	13/4.54	11/4.81	11/2.77	16/3.11	14/3.48	18/4.70	13.1/3.72	244/2.54	373/5.23	97/3.17	291/5.10	251.3/4.01	840/4.56
	RelFormer (Idan et al. 2024)	1/4.01	23/8.57	17/10.9	16/4.92	15/4.15	19/4.89	24/6.46	17.8/6.27	83/2.90	184/3.80	86/3.70	117/4.10	117.5/3.63	367/3.80
	DeMT (Ours)	8/3.59	7/3.53	4/2.60	9/3.81	8/3.55	9/3.51	6/2.19	7.3/3.25	23/1.19	18/1.02	9/0.87	12/0.98	15.5/1.02	24/0.94
RPR (mv)	CamNet <sup>†</sup> (Ding et al. 2019)	4/1.73	3/1.74	5/1.98	4/1.62	4/1.64	4/1.63	4/1.51	4.0/1.69	-/-	-/-	-/-	-/-	-/-	-/-
	RelPoseGNN (Turkoglu et al. 2021)	8/2.70	21/7.50	13/8.70	15/4.10	15/3.50	19/3.70	22/6.50	16.1/5.24	48/1.00	114/2.50	48/2.50	152/3.20	90.5/2.30	320/2.20
	ReLoc3r (Dong et al. 2025)	3/0.99	4/1.13	2/1.23	5/0.88	7/1.14	5/1.23	12/2.25	5.4/1.26	47/0.41	87/0.66	18/0.53	41/0.73	48.3/0.58	171/0.94
	DeMT (Ours)	2/0.74	3/0.95	2/1.18	4/0.72	5/0.97	4/1.16	4/1.80	3.6/1.12	9/0.54	8/0.56	5/0.37	6/0.44	7.0/0.48	11/0.39

Table 2: Median errors of baseline methods on benchmark VL datasets. Best RPR results are highlighted in bold and second best in underlined. Overall best results are marked in blue and second best are in green. “pw” and “mv” represent pair-wise and multi-view modes respectively. “-” means the value is unconverged or unobtainable. “†” means the method is unreproducible.

potential for MFVR. The integration of CMP enables multi-frame weighted fusion, achieving centimeter-level precision. Compared to competitive map-based methods such as DeViLoc, SACNet, and Maprepo, DeMT demonstrates superior flexibility and practicality. Notably, translation errors are consistently smaller in indoor environments than in outdoor settings, while rotation errors exhibit the opposite trend, aligning with their physical characteristics. Models tend to achieve higher accuracy on smaller scenes (e.g. Heads, Church) given the same training budget.

**MFVR Performance on Challenging Datasets.** Several competitive methods are further evaluated on more difficult datasets to assess their MFVR benefits, as detailed in Table 3. On the Aachen Day-Night v1.1 and Extended CMU-Seasons datasets, map-based methods (e.g. HLoc and DeViLoc) achieve top-tier outdoor localization accuracy. For instance, HLoc attains up to 95.4% accuracy on Aachen Day under the (0.5m, 2°) threshold. In contrast, DeMT delivers competitive accuracy with drastically lower computational cost, closely approaching the top-performing map-based methods. Crucially, it reduces model size by several orders of magnitude (e.g. 0.2GB vs. >15GB for DeViLoc) and cuts training time to less than 2.5 hours, while map-based pipelines often require days for map construction and optimization. These results, derived from InLoc and Aachen Day-Night v1.1 datasets featuring severe layout changes and illumination variations, underscore DeMT’s robustness and efficiency in complex real-world VL environments.

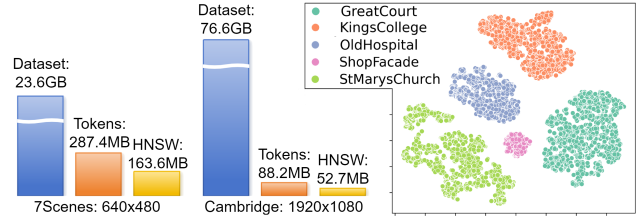


Figure 4: Tokenization results and t-SNE visualization.

**Scene Generalization.** The tokenization scheme endows MFVR with the unique capability to handle interleaved unseen scenes, as demonstrated in Table 4. DeMT outperforms direct regression methods, such as RelFormer, in scene identification tasks. However, RelPoseGNN exhibits significant performance degradation in outdoor environments due to increased false matches, while ReLoc3r fails entirely due to the absence of epipolar geometry constraints.

### Qualitative Visualization

The tokenization results with dimension 1536 are shown in Figure 4, while t-SNE (Maaten and Hinton 2008) also serves as a powerful tool to prove that DeMT produces more compact and separable feature clusters for outdoor scenes, indicating better scene generalization.

Furthermore, Figure 5 visualizes DeMT’s pipeline on five representative query images. These examples cover common

Dataset	Metrics (%) Method	InLoc (indoor, difficult)		Aachenv1.1 (outdoor, difficult)		Extend CMU-Seasons (outdoor, difficult)			Training Storage	
		Acc. (%) @ (0.25/0.5/1.0m, 10°) ↑		Acc. (%) @ (0.25/0.5/5.0m, 2/5/10°) ↑					Time ↓	Size ↓
		DUC1	DUC2	Day	Night	Urban	Suburban	Park	(hours)	(GB)
VPR	DenseVLAD (Jégou et al. 2010)	0.20/12.5/18.7	0.30/13.8/19.1	0.00/0.10/22.8	0.00/1.00/19.4	14.7/36.3/83.9	5.30/18.7/73.9	5.20/19.1/62.0	18	5.2
	NetVLAD (Arandjelovic et al. 2016)	8.20/24.7/48.9	6.40/26.3/54.9	0.00/0.20/18.9	0.00/0.00/14.3	12.2/31.5/89.8	3.70/13.9/74.7	2.60/10.4/55.9	30	4.8
	+Oracle (Sattler et al. 2019)	6.40/26.3/50.9	10.3/32.3/61.5	0.00/0.20/22.1	0.00/1.00/22.4	21.2/52.2/98.2	8.60/29.5/94.3	8.20/31.5/90.2	34	5.1
APE	AS (Sattler et al. 2017)	-/-/-	-/-/-	85.3/92.2/97.9	39.8/49.0/64.3	81.0/87.3/92.4	62.6/70.9/81.0	45.5/51.6/62.0	60	3.2
	D2Net (Dusmanu et al. 2019)	44.4/58.6/71.2	31.3/49.6/67.9	84.8/92.6/97.5	84.7/90.8/96.9	94.0/97.7/99.1	93.0/95.7/98.3	89.2/93.2/95.0	>99	22.3
	PixLoc (Sarlin et al. 2021)	25.5/47.3/68.8	32.4/54.7/79.5	74.3/79.3/87.4	61.0/65.8/79.3	88.3/90.4/93.7	79.6/81.1/85.2	61.0/62.5/69.4	>99	48.5
	HSCNet++(Wang et al. 2024b)	-/-/-	-/-/-	72.7/81.6/91.4	43.9/57.1/76.5	-/-/-	-/-/-	-/-/-	>99	27.4
	DeViLoc (Giang et al. 2024)	55.8/63.8/88.9	61.0/72.5/89.4	87.4/94.8/98.2	87.8/93.9/100.	95.7/98.4/99.2	97.1/98.3/99.4	92.1/95.1/96.3	>99	15.9
RPE	HLoc (SP+SG) (Sarlin et al. 2019)	49.0/68.7/80.8	53.4/77.1/82.4	89.6/95.4/98.8	86.7/93.9/100.	95.9/98.6/99.3	90.9/94.2/97.1	85.7/89.0/91.6	-	35.7
	RelFormer (Idan et al. 2024)	51.5/73.7/86.4	55.0/74.0/81.7	60.2/67.1/78.5	51.4/62.5/73.5	59.3/63.2/73.4	56.2/61.7/69.9	50.5/59.3/65.8	27	8.5
	ReLoc3r (Dong et al. 2025)	59.6/79.3/90.9	71.2/87.0/91.6	61.5/77.0/89.6	53.8/63.7/75.8	57.1/69.2/79.8	54.0/60.5/78.2	59.1/62.0/79.7	85	4.9
	DeMT (Ours)	64.9/84.8/91.3	75.6/88.3/92.8	88.7/95.0/98.3	86.2/91.8/99.5	87.5/95.9/97.6	82.5/89.6/93.8	83.7/88.9/90.5	≤2.5	0.2

Table 3: Average accuracies and training costs of competitive methods on challenging VL datasets. Overall best percentages are marked in blue and second best are in green. “-” means the value is not reported.

Scene	Method	7Scenes (m/°) ↓		Cambridge (m/°) ↓	
		Specific	Agnostic	Specific	Agnostic
pw	EssNet (Zhou et al. 2020b)	0.22/8.03	0.89/40.2	1.08/3.42	10.4/85.8
	NC-EssNet (Zhou et al. 2020b)	0.21/7.50	0.82/26.2	0.85/2.83	7.98/24.4
	Map-free (Arnold et al. 2022)	0.13/3.72	0.28/7.13	3.69/4.12	9.94/11.3
	RelFormer (Idan et al. 2024)	0.18/6.27	0.30/8.53	1.37/2.30	3.35/10.7
	DeMT (Ours)	<b>0.07/3.25</b>	<b>0.16/6.09</b>	<b>0.17/1.00</b>	<b>1.59/4.36</b>
mv	RelPoseGNN (Turkoglu et al. 2021)	0.16/5.24	0.36/13.6	1.68/3.60	-/-
	ReLoc3r (Dong et al. 2025)	-/-	<b>0.05/1.26</b>	-/-	0.73/0.65
	DeMT (Ours)	<b>0.03/1.12</b>	0.27/8.09	<b>0.08/0.46</b>	<b>0.59/0.72</b>

Table 4: Average errors of scene-specific/agnostic RPR. Best results are in bold and second best in underlined.

SKSA	TCDA	CMP	7Scenes (m/°) ↓		Cambridge (m/°) ↓	
✓	✓	✓	<b>0.073±0.017</b>	<b>3.25±0.56</b>	<b>0.173±0.061</b>	<b>1.00±0.11</b>
✓	✓	✓	0.118±0.032	5.53±0.85	0.196±0.096	1.99±0.64
✓	✗	✓	0.129±0.045	4.96±0.97	0.203±0.112	2.12±0.77
✓	✓	✗	0.153±0.086	7.82±1.65	0.281±0.177	3.47±1.08

Table 5: Average errors ± standard deviations by DeMT for ablation study on normal RPE datasets.

VL challenges such as viewpoint ambiguity, motion blur, illumination shifts, and dynamic occlusions. DeMT exhibits exceptional robustness across these diverse scenarios. Comparing columns (b) and (c), feature maps emphasize geometric details such as corners and edges, while heatmaps highlight semantic cues from discriminative landmarks (e.g. indoor signs and outdoor buildings), effectively avoiding distractions like closets or pedestrians that could cause mismatches. Comparing columns (d) and (e), the CMP demonstrates its efficacy in optimizing the selection of correct reference frames, particularly for orientation correction. Finally, we reconstruct a 3D grid map of each scene to calibrate these camera frustums and visualize their pose errors as shown in column (g).

## Ablation Study

**Component Ablations.** Leave-one-out variants are sequentially organized to evaluate the contribution of DeMT’s three innovative components: SKSA, TCDA, and CMP. As shown

Module	Technique	7Scenes (m/°) ↓		Cambridge (m/°) ↓		
Enc.	VIM	VGG16	0.129±0.026	5.96±0.79	0.396±0.102	3.12±0.77
		DINOv2-ViT	0.101±0.019	4.66±0.48	0.273±0.083	2.45±0.23
MIT	SKSA	NLSA	0.103±0.028	4.74±0.71	0.212±0.085	1.82±0.60
		w/o RBFCConv	0.097±0.021	4.15±0.66	0.195±0.071	1.46±0.29
		w/o Sparsemax	0.085±0.019	3.93±0.41	0.188±0.069	1.37±0.43
	TCDA	VLAD	0.144±0.033	7.49±1.15	0.744±0.781	4.69±1.72
GeM		0.156±0.029	7.84±0.91	0.971±0.265	4.64±1.09	
DAR	G	kNN	0.076±0.018	3.14±0.58	0.174±0.057	1.21±0.16
		Transformer Decoder	0.183±0.070	5.05±1.34	0.238±0.172	2.19±0.85
	CMP	Mamba Decoder	0.257±0.182	7.49±2.16	0.744±0.781	4.69±1.72
		Mamba Pointer	0.092±0.027	3.65±0.61	0.194±0.083	1.33±0.26
GPR	SSG	MLP	0.135±0.048	5.83±0.74	0.229±0.107	3.14±0.53
		w/o SwiGLU	0.121±0.031	4.60±0.59	0.214±0.096	2.91±0.68
		w/o Spectral	0.086±0.017	3.97±0.85	0.192±0.094	1.73±0.25
	DeMT (Ours)	<b>0.073±0.017</b>	<b>3.25±0.56</b>	<b>0.173±0.061</b>	<b>1.00±0.11</b>	

Table 6: Average errors ± standard deviations of different techniques for DeMT on normal RPE datasets.

in Table 5, TCDA imposes stricter constraints on MIT compared to SKSA, while CMP significantly enhances DAR by mitigating confounding factors between token proximity and pose similarity.

**Design Strategies.** Table 6 compares DeMT with other prominent techniques for MFVR. Its results affirm the sufficiency and necessity of DeMT’s design choices. Notably, the incremental HNSW algorithm outperforms the classic kNN algorithm in speed (approximately 5ms vs. 64ms per token) while delivering comparable accuracy. ViM proves to be more effective for visual encoding than VGG16 and DINOv2, consistent with the observation in Figure 6.

**Hyperparameter Analysis.** Figure 7 illustrates the performance of DeMT under varying hyperparameter configurations. It can be seen that the optimal accuracy is obtained with a feature dimension of 1536, 30 training epochs, and 10 refinement tokens.

## Latency and Memory

Table 7 outlines the computational overheads of various RPE models processing a query image in 640×480 resolu-

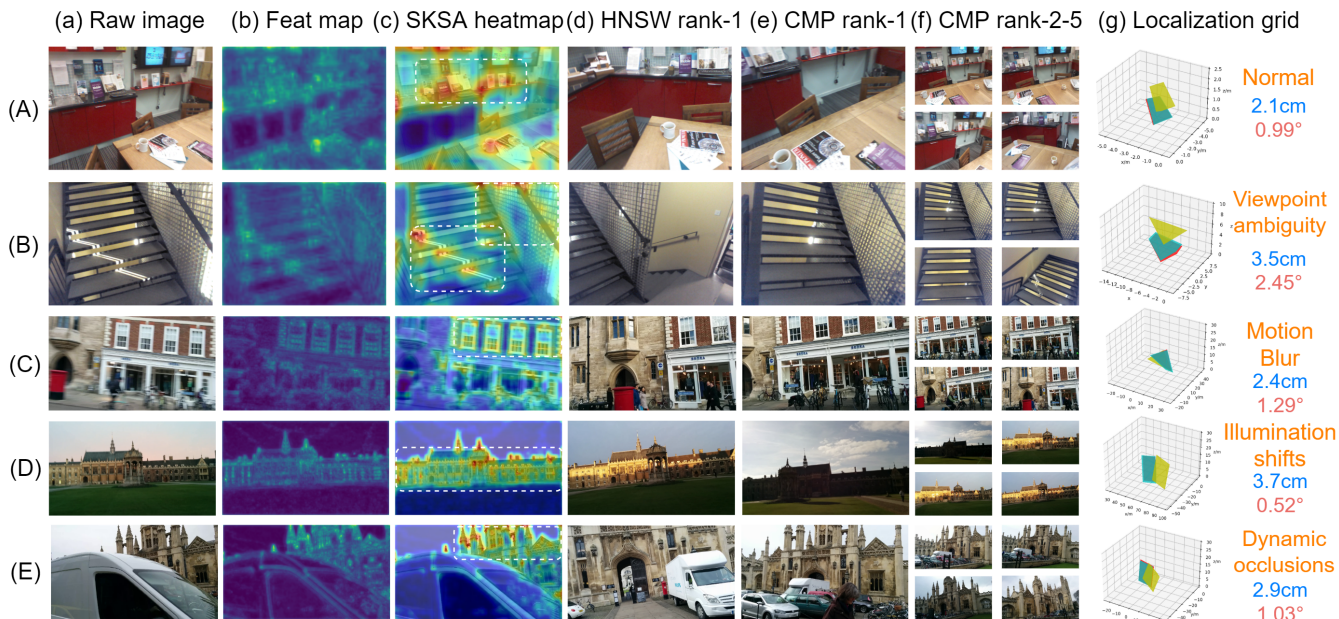


Figure 5: The visualization of DeMT for five examples: (a) raw images; (b) ViM feature maps; (c) attention maps of SKSA; (d) HNSW’s rank-1 images; (e) CMP’s rank-1 images; (f) CMP’s rank-2~5 images; (g) localization grids.

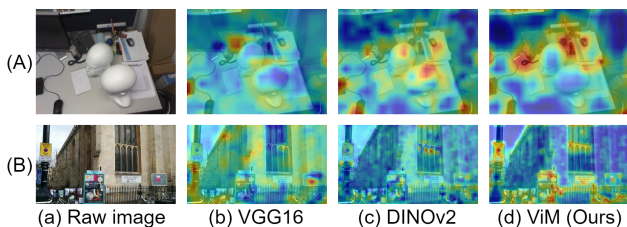


Figure 6: Attention heatmaps of CNN, ViT, and ViM.

Method	Extraction Latency (ms) ↓	Regression Latency (ms) ↓	Model Params (MB) ↓	Model FLOPs (GB) ↓
NN-Net (Laskar et al. 2017)	40.19	8.41	<b>47.32</b>	<b>87.6</b>
ReLocNet (Balntas et al. 2018)	41.80	9.32	75.30	116.4
Relformer (Idan et al. 2024)	106.47	18.92	291.45	359.8
Map-free (Arnold et al. 2022)	31.71	11.20	87.26	107.1
RelPoseGNN (Turkoglu et al. 2021)	72.63	28.28	204.41	284.2
Reloc3r (Dong et al. 2025)	83.09	135.36	353.89	419.8
DeMT (Ours)	<b>10.18</b>	<b>6.09</b>	269.17	274.5

Table 7: Latency and memory for online RPR inference on a single image. (FLOPs: floating point operations.)

tion. With superior indoor accuracy, DeMT is 3.11 and 8.16 times faster than Map-free and Reloc3r respectively in feature extraction. With the integration of SSG-MLPs for rapid convergence, DeMT reduces pose regression time to just 6.09ms. In contrast, ReLoc3r incurs a significant memory overhead exceeding 400GB FLOPs due to implicit stereo computations, whereas DeMT reduces this by 34.61%.

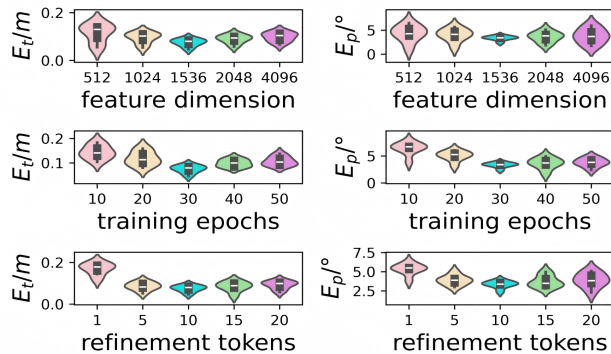


Figure 7: Parameter comparisons of DeMT: (A) feature dimension; (B) training epochs; (C) reference tokens. The left column is position errors, and the right is rotation errors.

## Conclusion

In this paper, we propose DeMT as an efficient and flexible solution for MFVR. By learning invariant scene features and employing causal reasoning over anchor tokens, DeMT effectively addresses the RPR problem in complex environments. Key innovations like CMP and SSG-MLP substantially reduce erroneous pose predictions. Extensive experiments validate DeMT’s superiority in terms of localization accuracy, inference speed, and scene generalization across diverse indoor and outdoor benchmarks. Nevertheless, its performance remains contingent on the representational quality of learned tokens and the relevance of retrieved reference images. Future work will target novel view synthesis to mitigate inherent multi-view redundancies.

## Acknowledgments

This work was jointly supported by the National Key Research and Development Program of China (No. 2024YFB4709802), National Natural Science Foundation of China (No. 62373009), Guangdong S&T Program (No. 2024B0101050002), the Major Key Project of Pengcheng Laboratory (No. PCL2024A01), and the Mobile Information Networks-National Science and Technology Major Project (No. 2025ZD1302900).

## References

- Ali-Bey, A.; Chaib-Draa, B.; Giguere, P.; et al. 2023. MixVPR: Feature mixing for visual place recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2998–3007.
- Amirian, M.; and Schwenker, F. 2020. Radial basis function networks for convolutional neural networks to learn similarity distance metric and improve interpretability. *IEEE Access*, 8: 123087–123097.
- Arandjelovic, R.; Gronat, P.; Torii, A.; et al. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5297–5307.
- Arnold, E.; Wynn, J.; Vicente, S.; et al. 2022. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision*, 690–708. Springer.
- Balntas, V.; Li, S.; Prisacariu, V.; et al. 2018. RelocNet: Continuous metric learning relocalisation using neural nets. In *European Conference on Computer Vision*, 751–767. Springer.
- Berton, G.; Masone, C.; Caputo, B.; et al. 2022. Rethinking visual geo-localization for large-scale applications. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4878–4888.
- Bjorck, N.; Gomes, C. P.; and Weinberger, K. Q. 2021. Towards deeper deep reinforcement learning with spectral normalization. *Advances in Neural Information Processing Systems*, 34: 8242–8255.
- Brachmann, E.; Cavallari, T.; Prisacariu, V. A.; et al. 2023. Accelerated coordinate encoding: Learning to relocalize in minutes using RGB and poses. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5044–5053.
- Brachmann, E.; and Rother, C. 2021. Visual camera relocalization from RGB and RGB-D images using DSAC. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 5847–5865.
- Bui, B.-T.; Bui, H.-H.; Tran, D.-T.; et al. 2024. D2S: Representing sparse descriptors and 3D coordinates for camera relocalization. *IEEE Robotics and Automation Letters*.
- Chen, Q.; Li, C.; Ning, J.; et al. 2024a. Gmconv: Modulating effective receptive fields for convolutional kernels. *IEEE transactions on neural networks and learning systems*, 36(4): 6669–6678.
- Chen, S.; Cavallari, T.; Prisacariu, V. A.; et al. 2024b. Map-relative pose regression for visual re-localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20665–20674.
- Chen, S.; Li, X.; Wang, Z.; et al. 2022. Dfnet: Enhance absolute pose regression with direct feature matching. In *European Conference on Computer Vision*, 1–17. Springer.
- Chi, L.; Jiang, B.; and Mu, Y. 2020. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33: 4479–4488.
- Ding, M.; Wang, Z.; Sun, J.; et al. 2019. CamNet: Coarse-to-fine retrieval for camera re-localization. In *IEEE International Conference on Computer Vision*, 2871–2880.
- Dong, S.; Wang, S.; Liu, S.; et al. 2025. Reloc3r: Large-Scale Training of Relative Camera Pose Regression for Generalizable, Fast, and Accurate Visual Localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Dusmanu, M.; Rocco, I.; Pajdla, T.; et al. 2019. D2-net: A trainable cnn for joint description and detection of local features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8092–8101.
- Giang, K. T.; Song, S.; Jo, S.; et al. 2024. Learning to produce semi-dense correspondences for visual localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19468–19478.
- Hatamizadeh, A.; and Kautz, J. 2025. Mambavision: A hybrid mamba-transformer vision backbone. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Hausler, S.; Garg, S.; Xu, M.; et al. 2021. Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14141–14152.
- Idan, O.; Shavit, Y.; Keller, Y.; et al. 2024. Beyond familiar landscapes: Exploring the limits of relative pose regressors in new environments. *SSRN preprint 5022843*, 1–9.
- Jégou, H.; Douze, M.; Schmid, C.; et al. 2010. Aggregating local descriptors into a compact image representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3304–3311.
- Kendall, A.; Grimes, M.; Cipolla, R.; et al. 2015. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In *IEEE International Conference on Computer Vision*, 2938–2946.
- Laskar, Z.; Melekhov, I.; Kalia, S.; et al. 2017. Camera relocalization by computing pairwise relative poses using convolutional neural network. In *IEEE International Conference on Computer Vision Workshops*, 929–938.
- Lin, J.; Gu, J.; Wu, B.; et al. 2024. Learning neural volumetric pose features for camera localization. In *European Conference on Computer Vision*, 198–214. Springer.
- Liu, X.; and Zhang, Y. 2023. Matrices over quaternion algebras. In *Matrix and Operator Equations and Applications*, 139–183. Springer.
- Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11): 2579–2605.

- Malkov, Y. A.; and Yashunin, D. A. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4): 824–836.
- Martins, A.; and Astudillo, R. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, 1614–1623. PMLR.
- Moreau, A.; Piasco, N.; Tsishkou, D.; et al. 2022. Lens: Localization enhanced by nerf synthesis. In *Conference on Robot Learning*, 1347–1356. PMLR.
- Oquab, M.; Darcet, T.; Moutakanni, T.; et al. 2023. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv: 2304.07193*.
- Polizzi, V.; Cannici, M.; Scaramuzza, D.; et al. 2025. FaVoR: Features via voxel rendering for camera relocalization. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 44–53.
- Saha, S.; Varma, G.; Jawahar, C. V.; et al. 2018. Improved visual relocalization by discovering anchor points. In *British Machine Vision Conference*.
- Sarlin, P.-E.; Cadena, C.; Siegwart, R.; et al. 2019. From coarse to fine: Robust hierarchical localization at large scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12716–12725.
- Sarlin, P.-E.; Unagar, A.; Larsson, M.; et al. 2021. Back to the feature: Learning robust camera localization from pixels to pose. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3247–3257.
- Sattler, T.; Leibe, B.; Kobbelt, L.; et al. 2017. Efficient and effective prioritized matching for large-scale image-based localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9): 1744–1756.
- Sattler, T.; Zhou, Q.; Pollefeys, M.; et al. 2019. Understanding the limitations of CNN-based absolute camera pose regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3297–3307.
- Shavit, Y.; and Keller, Y. 2022. Camera pose auto-encoders for improving pose regression. In *European Conference on Computer Vision*, 140–157. Springer.
- Shazeer, N. 2020. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*.
- Shotton, J.; Glocker, B.; Zach, C.; et al. 2013. Scene coordinate regression forests for camera relocalization in RGB-D images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2930–2937.
- Taira, H.; Okutomi, M.; Sattler, T.; et al. 2018. InLoc: Indoor visual localization with dense matching and view synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7199–7209.
- Tang, S.; Tang, S.; Tagliasacchi, A.; et al. 2023. Neumap: Neural coordinate mapping by auto-transdecoder for camera localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 929–939.
- Toft, C.; Maddern, W.; Torii, A.; et al. 2020. Long-term visual localization revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4): 2074–2088.
- Turkoglu, M. O.; Eric, B.; Konrad, S.; et al. 2021. Visual camera re-localization using graph neural networks and relative pose supervision. In *IEEE International Conference on 3D Vision*, 145–155.
- Wang, K.; Jiang, Z.; Dai, K.; et al. 2024a. SACNet: A scattered attention-based network with feature compensator for visual localization. *IEEE Robotics and Automation Letters*, 9(4): 3586–3593.
- Wang, R.; Shen, Y.; Zuo, W.; et al. 2022. TransVPR: Transformer-based place recognition with multi-level attention aggregation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13648–13657.
- Wang, S.; Laskar, Z.; Melekhov, I.; et al. 2024b. HSCNet++: Hierarchical scene coordinate classification and regression for visual localization with transformer. *International Journal of Computer Vision*, 132(7): 2530–2550.
- Wang, Z.; Xiang, X.; Zhao, Z.; and Su, F. 2018. Deep image retrieval: Indicator and gram matrix weighting for aggregated convolutional features. In *IEEE International Conference on Multimedia and Expo*, 1–6.
- Yang, M.; He, D.; Fan, M.; et al. 2021. DOLG: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *IEEE International Conference on Computer Vision*, 11772–11781.
- Yeh, C.; Chen, Y.; Wu, A.; et al. 2024. Attentionviz: A global view of transformer attention. *IEEE Transactions on Visualization and Computer Graphics*, 30(1): 262–272.
- Zeng, K.; Lin, H.; Yan, Z.; et al. 2024. Non-local self-attention network for image super-resolution. *Applied Intelligence*, 54(7): 5336–5352.
- Zhang, H.; Zhu, Y.; Wang, D.; et al. 2024. A survey on visual mamba. *Applied Sciences*, 14(13): 5683.
- Zhou, D.; Hou, Q.; Chen, Y.; et al. 2020a. Rethinking bottleneck structure for efficient mobile network design. In *European Conference on Computer Vision*, 680–697. Springer.
- Zhou, Q.; Sattler, T.; Pollefeys, M.; et al. 2020b. To learn or not to learn: Visual localization from essential matrices. In *IEEE International Conference on Robotics and Automation*, 3319–3326.