Debiased Multiplex Tokenizer for Efficient Map-Free Visual Relocalization

Wenshuai Wang[1,2], **Hong Liu**[1,*], **Shengquan Li**[2], **Peifeng Jiang**[1], **Runwei Ding**[2,*]

[1]State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School
[2]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 position and orientation 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 nine 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 geotagged 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 relocalization (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 committed 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.

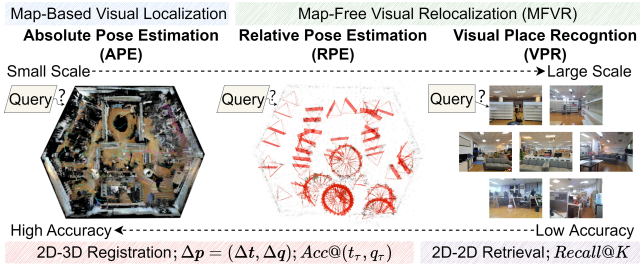

Figure 1: Main paradigms of visual localization. For a query image, absolute pose estimation (APE) matches it directly to the 3D scene map, while relative pose estimation (RPE) and visual place recognition (VPR) need to retrieve the reference dataset. Moreover, RPE is supposed to make a better trade-off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

*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. Besides, 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 a causal reasoning to pose correlation. It not only imposes frequency calibration into feature description to overcome potential interferences from motion blurs or illumination changes, but also achieves linear time complexity via the coupled 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 datasets, exemplified by indoor 7Scenes (Shotton et al. 2013) and outdoor Cambridge Landmarks (Kendall et al. 2015) datasets. In particular, all angular values are unified into unit quaternions to prevent the gimbal lock problem (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-MLP ensure accurate and interpretable 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.

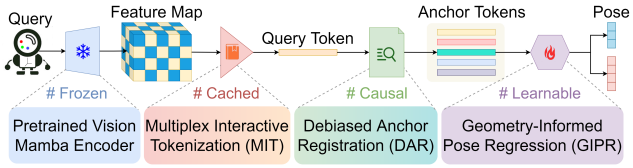

Figure 2: The pipeline of our proposed method. Four modules are cascaded for relative pose regression (RPR).

Related Work

Visual Place Recognition

VPR is typically formulated as an image retrieval (IR) problem governed by similarity thresholds. NetVLAD (Arandjelovic et al. 2016) integrates a CNN with the vector of local aggregated descriptors (VLAD), establishing a pioneering approach for global feature aggregation. DOLG (Yang et al. 2021) employs orthogonal projection to decompose local descriptors and explicitly concatenates them with global features. Patch-NetVLAD (Hausler et al. 2021) and TransVPR (Wang et al. 2022) leverage ViTs to construct feature pyramids for patch-level fusion. To mitigate redundancy in explicit methods, implicit techniques have gained traction. CosPlace (Berton et al. 2022) adopts generalized mean pooling (GeM) for domain adaptation, while MixVPR (Ali-Bey et al. 2023) enhances context mining through MLP-based mixer blocks. SelaVPR (Lu et al. 2024) fine-tunes pre-trained DINOv2 (Oquab et al. 2023) features using joint adaptation of global and local feature, enabling more effective assignment of static landmarks.

Absolute Pose Estimation

APE aims to establish precise stereo 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 advanced scene coordinate regression (SCR) and absolute pose regression (APR), represented by the DSAC family (Brachmann and Rother 2021) and PoseNet family (Kendall et al. 2015), respectively. However, their effectiveness in dynamic environments is limited by mapping costs and the requirement for accurate pose regression (Chen et al. 2024b). SAC-Net (Wang et al. 2024a), LENS (Moreau et al. 2022), and Maprepo (Chen et al. 2024b) focus on feature compression, synthetic dataset 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 (Sarlin 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 using anchor frame clusters. NN-Net (Laskar et al. 2017) examines pairwise differences via Siamese networks, while ReLocNet (Balntas

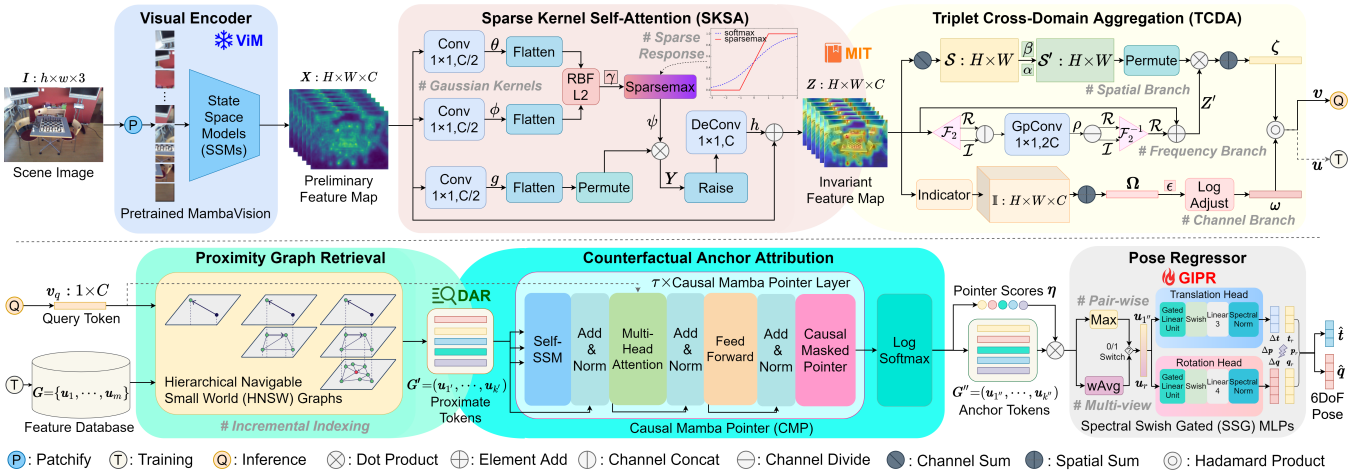

Figure 3: The overview of Debiased Multiplex Tokenizer (DeMT). This framework sequentially performs self-attentive mapping from images to tokens and causal inference to poses. RBF, DeConv, and GpConv stand for radial basis function, deconvolution, and grouped convolution, respectively.

et al. 2018) and EssNet (Zhou et al. 2020b) enhance its accuracy through continual learning and geometric constraints, respectively. RelFormer (Idan et al. 2024) employs a dual-branch Transformer to address scene generalization challenges. RelPoseGNN (Turkoglu 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) RPR performance on the indoor 7Scenes dataset but generalizes poorly to outdoor environments.

Methodology

Feature Map Encoding

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

Multiplex Interactive Tokenization

MIT aggregates $\boldsymbol{X}$ into a unique token $\boldsymbol{v} \in \mathbb{R}^C$ through two 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. We replace the dot product with exponential Euclidean distance to capture invariant non-local responses in the Gaussian kernel space, defined as:

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

where $\theta(\cdot), \phi(\cdot)$ are affinity kernels, and $\boldsymbol{W}$ corresponds to their learnable weights. Actually, $f_{\text{gk}}(\cdot)$ is equivalent to a radial basis function (RBF) with scaling factor $\gamma > 0$.

Sparse Response. To emphasize invariant feature affinity, sparsemax (Martins and Astudillo 2016) is used to upgrade conventional softmax and prevent overfitting, yielding the SKSA response $\boldsymbol{Y} \in \mathbb{R}^{HW\times C}$:

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

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

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

$$\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{Y}\boldsymbol{W}_h. \tag{3}$$

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

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

$$\boldsymbol{S'} = [\frac{\boldsymbol{S}}{(\boldsymbol{J}_{W,H}\boldsymbol{S}^\alpha)^{\frac{1}{\alpha}}}]^{\frac{1}{\beta}}, \tag{4}$$

where α and β are scaling factors.

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

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

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; $\boldsymbol{W}_\rho$ is the weight matrix of 1×1 grouped convolution.

Interim weighted pooling is then performed to yield the spatial-frequency token $\boldsymbol{\zeta} \in \mathbb{R}^C$ as $\boldsymbol{\zeta} = \boldsymbol{J}_{W,H}\boldsymbol{S'}^\top \boldsymbol{Z'}$.

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

$$\boldsymbol{\Omega} = \frac{1}{HW}\boldsymbol{J}_{W,H}\mathbb{I}(\boldsymbol{Z} > \boldsymbol{O}), \tag{6}$$

where $\mathbb{I}(\cdot)$ is an indicator function and $\boldsymbol{O}$ is an all-zero tenser.

This is logarithmically adjusted to $\boldsymbol{\omega} \in \mathbb{R}^C$ with a stability constant ϵ:

$$\boldsymbol{\omega} = \log(\frac{C\epsilon + \boldsymbol{\Omega}\boldsymbol{J}_{C,1}}{\epsilon + \boldsymbol{\Omega}}). \tag{7}$$

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

Debiased Anchor Registration

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

Proximity Graph Retrieval. Given a query token $\boldsymbol{v}_q$ and reference token database represented by proximity graphs $\boldsymbol{G} = \{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$, the HNSW algorithm $f_{\text{hnsw}}(\cdot,\cdot)$ incrementally retrieves the top-k proximate candidates:

$$\boldsymbol{G'} = f_{\text{hnsw}}(\boldsymbol{D}, \boldsymbol{v}_q) = (\boldsymbol{u}_{1'}, \cdots, \boldsymbol{u}_{k'}), \tag{8}$$

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

Counterfactual Anchor Attribution. To validate feature-pose correspondences, a τ-layer CMP $f^{\tau}_{\text{cmp}}(\cdot,\cdot)$ debiases $\boldsymbol{P}$ into anchor set $\boldsymbol{A}$:

$$\boldsymbol{G''} = f^{\tau}_{\text{cmp}}(\boldsymbol{G}, \boldsymbol{v}_q) = (\boldsymbol{u}_{1''}, \cdots, \boldsymbol{u}_{k''}), \tag{9}$$

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 $\Delta\boldsymbol{p} = [\Delta\boldsymbol{t} \in \mathbb{R}^3, \Delta\boldsymbol{q} \in \mathbb{R}^4]$ as:

$$\Delta\boldsymbol{p} = [f_{sn}(f_{sw}(\boldsymbol{t}_q) \odot \sigma(\boldsymbol{t}_q)), f_{sn}(f_{sw}(\boldsymbol{q}_q) \odot \sigma(\boldsymbol{q}_q))], \tag{10}$$

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

Therefore, for a reference token $\boldsymbol{v}_r$ (with labeled pose $\boldsymbol{p}_r = [\boldsymbol{t}_r, \boldsymbol{q}_r]$), the predicted pose $\hat{\boldsymbol{p}}^{(r)}_q = [\hat{\boldsymbol{t}}^{(r)}_q, \hat{\boldsymbol{q}}^{(r)}_q]$ is calculated as:

$$\hat{\boldsymbol{p}}^{(r)}_q = [\boldsymbol{t}_r + \Delta\boldsymbol{t}, f_{\text{quat}}(f_{\text{rot}}(\boldsymbol{q}_r) f_{\text{rot}}(\Delta\boldsymbol{q}))], \tag{11}$$

where $f_{\text{rot}}(\cdot)$ and $f_{\text{quat}}(\cdot)$ represent mutual conversions between unit quaternions and rotation matrices.

Loss and Inference

Training Loss. We can define the RPR loss L_{rpr} as pose differences by using L_1 and L_2 norms:

$$L_{\text{rpr}} = ||\boldsymbol{t}_q - \hat{\boldsymbol{t}}_q||_1 + ||\boldsymbol{q}_q - \hat{\boldsymbol{q}}^{(r)}_q / ||\hat{\boldsymbol{q}}^{(r)}_q||_2||_1. \tag{12}$$

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

$$L_{\text{total}} = L_{\text{rpr}} + L_{\text{hnsw}} + L_{\text{cmp}}, \tag{13}$$

where L_{hnsw} and L_{cmp} are negative log-likelihood (NLL) losses for distribution differences from $\boldsymbol{G'}$ and $\boldsymbol{G''}$ to $\boldsymbol{v}_q$'s groundtruth respectively.

Table 1: Details of the adopted APE/RPE and VPR datasets.

	Dataset	# refer.	# query	Motion	Light	Season	Occlusion
APE/RPE	7Scenes (Shotton et al. 2013)	26.0k	17.0k	★★☆	★☆☆	☆☆☆	☆☆☆
APE/RPE	Cambridge (Kendall et al. 2015)	8,380	4,841	★★★	★★☆	☆☆☆	★★☆
APE/RPE	InLoc (Taira et al. 2018)	9,972	329	★★☆	★★☆	☆☆☆	★★★
APE/RPE	Aachenv1.1 (Toft et al. 2020)	6,697	1,015	★★☆	★★★	☆☆☆	★★★
APE/RPE	Extend CMU-Seasons (Toft et al. 2020)	60.9k	56.6k	★★★	★★☆	★★★	★★☆
VPR	Pitts250k-test (Arandjelovic et al. 2016)	83.9k	8.2k	★★☆	★★☆	★☆☆	★★☆
VPR	MSLS-val (Warburg et al. 2020)	18.9k	740	★★★	★☆☆	★★☆	★☆☆
VPR	MSLS-chall. (Warburg et al. 2020)	38.8k	27.1k	★★☆	★★★	★★★	★★★
VPR	Nordland-test (Sünderhauf et al. 2013)	27.6k	3.5k	☆☆☆	★☆☆	★★★	☆☆☆

Online Inference. DeMT supports for two RPR modes: pair-wise calibration and multi-view collaboration.

Pair-wise Mode. Only the top-ranked anchor $\boldsymbol{u}_{1''}$ is selected from $\boldsymbol{G''}$, so the absolute pose error is computed as $E_{pw}(\boldsymbol{v}_q, \boldsymbol{u}_{1''}) = (||\boldsymbol{t} - \hat{\boldsymbol{t}}||_2,\ 2\arccos|\boldsymbol{q}\hat{\boldsymbol{q}}|\frac{180^\circ}{\pi})$.

Multi-view Mode. Due to the Lipschitz continuity of spectral normalization (Bjorck, Gomes, and Weinberger 2021), a structural smoothing error can be obtained by directly weighting average on multiple frames with their confidence scores $\boldsymbol{\eta} \in [0,1]^k$ from CMP as $E^k_{mv}(\boldsymbol{v}_q, \boldsymbol{G''}) = \sum_{i=1}^{k} \eta_i E_{pw}(\boldsymbol{v}_q, \boldsymbol{u}_{i''})$.

Experiments

Datasets and Evaluation Metrics

Datasets. DeMT is evaluated on five APE/RPE datasets and four VPR datasets as detailed in Table 1. APE/RPE datasets include indoor and outdoor types, as well as benchmark and challenge instances, while VPR datasets are mainly captured from large-scale streetscapes. All datasets are publicly available to ensure experimental reproducibility.

Evaluation Metrics. APE/RPE performance is assessed by absolute pose errors E_{pw} or E_{mv}. To facilitate an intuitive evaluation, we also report their median, average, standard deviation, and specific accuracy. VPR performance is quantified by Recall@1/5/10 that calculated as the ratio of correct matches among the top-ranked reference images.

Implementation Details

Models. The pretrained MbambaVision-L (Hatamizadeh and Kautz 2025) is choosen as the visual encoder with fixed feature channel dimension C=1536. The final average pooling and fully connected layers are excluded from this architecture for subsequent tokenization.

Hyperparameters. For MIT, $\theta(\cdot)$, $\phi(\cdot)$ in Equation (1) are implemented by 1×1 radial basis function (RBF) convolution (Amirian and Schwenker 2020), while $g(\cdot)$ is by 1×1 bottleneck convolution (Zhou et al. 2020a). α and β in Equation (4) are set to 0.5 and 2, respectively. ϵ in Equation (7) is set to 0.0001. For DAR, HNSW's k is set to 10, and CMP's τ 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

Table 2: Median errors of baseline methods on benchmark VL datasets, while accuracies at specific thresholds are partially provided. Best RPR results are highlighted in **bold** and second best in underlined. Overall best and second best results are marked in blue and green. "pw" and "mv" represent pair-wise and multi-view modes respectively. "–" means the value is unconverged or unobtainable. "†" means the method is unreproducible to our effort.

Dataset	Metrics / Method	7Scenes (indoor, standard) Med. Err. (cm/°) ↓ | Acc. (%)@(5cm, 5°) ↑								Cambridge Landmarks (outdoor, standard) Med. Err. (cm/°) ↓ | Acc. (%)@(0.5m, 5°) ↑					
		Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Average	College	Hospital	Shop	Church	Average-4	Court
SbP	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$^{68.7}$	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$^{75.7}$	14/0.24	16/0.32	5/0.23	10/0.34	11.3/0.28	30/0.14
	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
	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 and Rother 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$^{88.7}$	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	2.7/0.96$^{80.8}$	28/0.40	31/0.60	5/0.30	18/0.60	20.5/0.48	43/0.20
	HSCNet++(Wang et al. 2024b)	2/0.70$^{98.1}$	2/0.72$^{97.0}$	1/0.80$^{98.8}$	2/0.69$^{88.2}$	4/1.00$^{65.1}$	4/1.15$^{72.9}$	3/1.02$^{76.6}$	2.6/1.36$^{85.2}$	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$^{98.6}$	2/0.74$^{92.9}$	1/0.75$^{98.7}$	2/0.62$^{91.3}$	3/0.83$^{72.8}$	3/1.04$^{77.9}$	13/2.02$^{24.2}$	3.7/0.94$^{79.5}$	7/0.12	15/0.29	3/0.17	8/0.25	8.3/0.21	23/0.11
	SACNet (Wang et al. 2024a)	2/0.53$^{97.1}$	2/0.71$^{92.9}$	1/0.59$^{99.8}$	3/0.65$^{89.3}$	2/0.65$^{82.0}$	3/0.93$^{80.5}$	2/0.40$^{87.0}$	2.1/0.64$^{89.8}$	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† (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† (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	246/4.01	840/4.56
	Relformer (Idan et al. 2024)	11/4.01	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$^{18.5}$	**7/3.53**$^{23.2}$	**4/2.60**$^{62.1}$	9/3.81$^{17.3}$	**8**/3.55$^{18.4}$	**9/3.51**$^{13.0}$	**6/2.19**$^{37.8}$	**7.3/3.25**$^{27.2}$	**23**/1.19$^{92.5}$	**18/1.02**$^{94.5}$	**9/0.87**$^{97.3}$	**12/0.98**$^{98.6}$	**15.5/1.02**$^{96.0}$	**24/0.94**$^{97.3}$
RPR (mv)	CamNet† (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**$^{98.3}$	**3/0.95**^{100}	**2/1.18**^{100}	**4/0.72**$^{97.5}$	**5/0.97**$^{98.1}$	**4/1.16**$^{97.8}$	**4/1.80**$^{99.5}$	**3.6/1.12**$^{98.7}$	9/0.54$^{96.5}$	**8/0.56**$^{98.4}$	**5/0.37**$^{99.0}$	**6/0.44**^{100}	**7.0/0.48**$^{98.7}$	**11/0.39**$^{99.8}$

Table 3: Average accuracies of competitive methods on challenging VL datasets. Best/second best RPE results are highlighted in **bold**/underlined. Overall best/second best results are marked in blue/green. Training time and mapping overhead are sampled.

Dataset	Metrics (%) / Method	InLoc (indoor, difficult) Acc. (%) @(0.25/0.5/1.0m, 10°) ↑		Aachenv1.1 (outdoor, difficult) Acc. (%) @(0.25/0.5/5.0m, 2/5/10°) ↑		Extend CMU-Seasons (outdoor, difficult) Acc. (%) @(0.25/0.5/5.0m, 2/5/10°) ↑			Training Time ↓ (hours)	Map/Token Size ↓ (GB)
		DUC1	DUC2	Day	Night	Urban	Suburban	Park		
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	36	5.1
APE	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	≥168	48.5
	DFNet (Chen et al. 2022)	39.3/63.4/83.2	45.5/68.6/87.8	80.5/87.4/94.2	68.4/77.6/88.8	63.1/66.9/80.5	62.4/65.9/78.3	58.7/62.3/75.8	≥168	22.3
	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	≥168	15.9
	SACNet (Wang et al. 2024a)	34.3/61.9/79.2	45.5/67.8/83.9	65.5/77.3/88.8	52.3/68.7/79.1	65.5/72.3/88.8	58.4/67.5/80.3	54.3/69.3/76.4	60	3.12
RPE	HFNet (Sarlin et al. 2019)	49.3/68.5/80.8	54.1/76.9/82.4	75.7/84.3/90.9	40.8/55.1/72.4	80.4/83.5/91.6	71.8/78.2/87.1	67.4/75.3/90.4	–	35.7
	Map-free (Arnold et al. 2022)	47.0/71.2/84.8	58.8/77.9/80.9	79.1/84.7/89.4	70.2/81.8/91.3	78.4/87.7/92.5	71.2/76.5/79.3	56.0/59.5/62.1	84	0.8
	RelFormer (Idan et al. 2024)	53.5/76.8/85.9	61.8/80.9/87.0	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	≥168	8.5
	Reloc3r (Dong et al. 2025)	59.6/79.3/89.3	65.2/83.6/90.7	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	≥168	1.8
	DeMT (Ours)	**64.9/84.8/91.4**	**74.6/86.3/92.8**	**82.7/90.9/95.3**	**76.2/78.5/79.9**	**79.5/87.9/97.6**	**72.5/88.8/92.8**	**77.3/82.8/87.5**	**≤2.5**	**0.2**

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 presents a comprehensive evaluation on RPR performance of DeMT, alongside representative VL methods 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 reveals underscore DeMT's significant potential for MFVR. Additionally, 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 responsiveness. Notably, translation errors are consistently smaller in indoor environments than in outdoor settings, while rotation errors exhibit the opposite trend, aligning with physical expectations. Smaller spaces and fewer images facilitate higher accuracy after equivalent training epochs, as evidenced by Heads and Church scenes.

MFVR Performance on Challenging Datasets. Several competitive methods are further investigated for their MFVR benefits on more difficult datasets as detailed in Table 3. Although DeViLoc outperforms DeMT by 8.9% on average, DeMT is the most lightweight competitor. The results are derived from the InLoc and Aachen datasets, which feature significant challenges such as drastic layout changes and varying light conditions, underscoring DeMT's robustness in complex real-world environments.

VPR Performance. As shown in Table 4, DeMT achieves average improvements in Recall@1/5/10 of 0.6/0.8/0.4%, respectively on MSLS-val datasets compared to SelaVPR descriptors. These results validate the efficacy of our design, particularly when compared to methods rely on DINOv2-ViTs with quadratic linear complexity. Furthermore, the integration of local feature mixtures significantly enhances re-

Table 4: Recall@1/5/10 comparisons on VPR datasets. Best results are in **bold** and second best in underlined.

Method	Pitts250k-test R@1/5/10 (%) ↑	MSLS-val R@1/5/10 (%) ↑	MSLS-challenge R@1/5/10 (%) ↑	Nordland-test R@1/5/10 (%) ↑
NetVLAD (Arandjelovic et al. 2016)	81.9/91.2/93.7	52.4/64.7/69.4	31.5/42.1/46.2	10.9/19.2/24.5
DOLG (Yang et al. 2021)	89.9/95.4/96.7	82.0/88.9/91.4	75.6/87.1/90.8	51.3/66.8/69.8
CosPlace (Berton et al. 2022)	88.4/94.5/95.7	82.8/89.7/92.0	61.4/72.0/76.6	54.4/69.8/75.9
Conv-AP (Ali-bey et al. 2022)	92.4/95.7/98.4	83.4/90.5/92.3	75.0/86.8/90.3	38.2/54.8/61.2
Patch-NV (Hausler et al. 2021)	88.7/94.5/95.9	79.5/86.2/87.7	48.1/57.6/60.5	51.6/60.1/62.8
TransVPR (Wang et al. 2022)	89.0/94.9/96.2	86.8/91.2/92.4	63.9/74.0/77.5	61.3/71.7/75.6
MixVPR (Ali-Bey et al. 2023)	91.5/95.5/96.3	88.0/92.7/94.6	64.0/75.9/80.6	58.4/74.6/80.0
SelaVPR (Lu et al. 2024)	92.7/98.0/98.9	87.7/95.8/96.6	69.6/86.9/90.1	47.2/66.6/74.1
DeMT (Ours)	**93.5/98.3/99.2**	**88.3/96.6/97.0**	**76.4/87.7/91.6**	**63.7/79.3/84.8**

Table 5: Average errors of scene-specific/agnostic RPR. Best results are in **bold** and second best in underlined.

Scene	Method	7Scenes (m/ °) ↓ Specific	7Scenes (m/ °) ↓ Agnostic	Cambridge (m/ °) ↓ Specific	Cambridge (m/ °) ↓ 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)	**0.07/3.25**	**0.16/6.09**	**0.17/1.00**	**1.59/4.36**
mv	RelPoseGNN (Turkoglu et al. 2021)	0.16/5.24	0.36/13.6	1.68/3.60	– / –
	Reloc3r (Dong et al. 2025)	– / –	**0.05/1.26**	– / –	0.73/**0.65**
	DeMT (Ours)	**0.03/1.12**	0.27/8.09	**0.08/0.46**	**0.59**/0.72

trieval performance, surpassing explicit methods like DOLG and implicit approaches like MixVPR in challenging cases.

Scene Generalization. Our image tokenization strategy endows MFVR with the unique capability to handle interleaved unseen scenes, as demonstrated in Table 5. DeMT outperforms direct regression methods, such as Relformer, in scene identification tasks. However, ReLocGNN 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

Figure 4 illustrates the visualization results of DeMT across four query images, representing VL challenges such as perspective perturbations, repeated textures, lighting variations, and pedestrian movements, respectively. 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 voxel map of each scene to calibrate these camera frustums and visualize their pose errors as shown in column (g).

Furthermore, the tokenization results with dimension 1536 are shown in Figure 5, while t-SNE also serves as a powerful tool for feature aggregation to prove DeMT is blessed with better Kullback-Leibler divergences for scene generalization on Cambridge Landmarks dataset.

Ablation Study

Component Ablation. Leave-one-out variants are organized to evaluate the contribution of DeMT's three innovative components: SKSA, TCDA, and CMP. As shown in Table 6, TCDA imposes stricter constraints on MIT compared to SKSA, while CMP significantly enhances DAR by mitigating unknown confounders between proximate tokens and similar poses.

Table 6: Average errors ± standard deviations by DeMT for ablation study on normal RPE datasets.

SKSA	TCDA	CMP	7Scenes (m/ °) ↓	Cambridge (m/ °) ↓
✔	✔	✔	**0.073±0.017/3.25±0.56**	**0.173±0.061/1.00±0.11**
✘	✔	✔	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 7: Average errors ± standard deviations of different techniques for DeMT on normal RPE datasets.

	Technique	7Scenes (m/ °) ↓	Cambridge (m/ °) ↓
Enc.	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
	MambaVision (Ours)	**0.073±0.017/3.34±0.43**	**0.173±0.061/1.00±0.11**
MIT	NLSA	0.103±0.028/4.74±0.71	0.162±0.075/1.82±0.60
	w/o RBFConv	0.097±0.021/4.15±0.66	0.155±0.071/1.46±0.29
	w/o Sparsemax	0.085±0.019/3.93±0.41	0.138±0.059/1.37±0.43
	SKSA (Ours)	**0.073±0.017/3.34±0.43**	**0.173±0.061/1.00±0.11**
	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±0.267/4.64±1.09
	w/o Frequency	0.096±0.019/4.59±0.63	0.207±0.021/1.74±0.63
	TCDA (Ours)	**0.073±0.017/3.34±0.43**	**0.173±0.061/1.00±0.11**
DAR	kNN	0.076±0.018/3.48±0.47	0.174±0.066/1.21±0.16
	HNSW (Ours)	**0.073±0.017/3.34±0.43**	**0.173±0.061/1.00±0.11**
	Transformer Decoder	0.183±0.070/5.05±1.34	0.238±0.172/2.19±0.85
	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±0.60	0.194±0.083/1.33±0.26
	CMP (Ours)	**0.073±0.017/3.34±0.43**	**0.173±0.061/1.00±0.11**
GIPR	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
	SSG-MLP (Ours)	**0.073±0.017/3.34±0.43**	**0.173±0.061/1.00±0.11**

Design Strategies. Table 7 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 is proved to be much effective for visual encoding compared to VGG16 and DINOv2, which is also illustrated in Figure 6.

Hyperparameter Analysis. Figure 7 reveals that DeMT achieves optimal performance with feature dimension 1536, 30 training epochs, 10 reference tokens.

Table 8: Latency and memory for online RPR inference on a single image. (FLOPs: floating point operations.)

Method	Extraction time (ms)	Regression time (ms)	Params (MB)	FLOPs (GB)
NN-Net (Laskar et al. 2017)	40.19	8.41	**23.66**	**44.29**
ReLocNet (Balntas et al. 2018)	41.80	9.32	25.10	46.50
Relformer (Idan et al. 2024)	106.47	18.92	291.42	359.90
Map-free (Arnold et al. 2022)	31.71	11.20	187.22	107.01
RelPoseGNN (Turkoglu et al. 2021)	72.63	28.28	204.41	184.42
Reloc3r (Dong et al. 2025)	83.09	235.36	353.89	274.53
DeMT-pw (Ours)	**10.18**	**6.09**	264.95	119.85

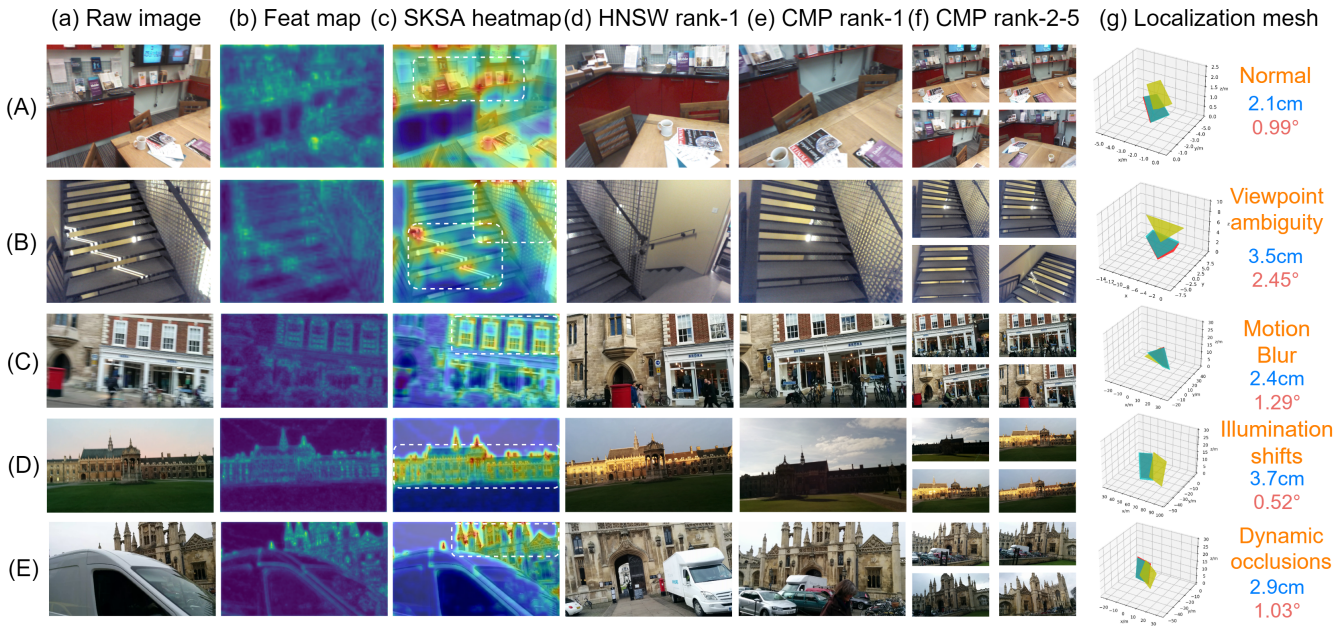

Figure 4: The visualization of DeMT for five examples: (a) raw images; (b) ViM feature maps; (c) attention maps of SKSA; (d) HNSW's rank-1 images; (e) CMP's rank-1 images; (f) CMP's rank-2∼5 images; (g) localization meshes.

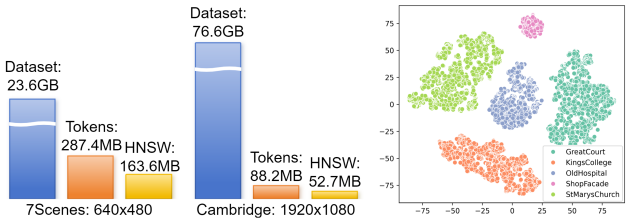

Figure 5: Tokenization results and t-SNE visualization.

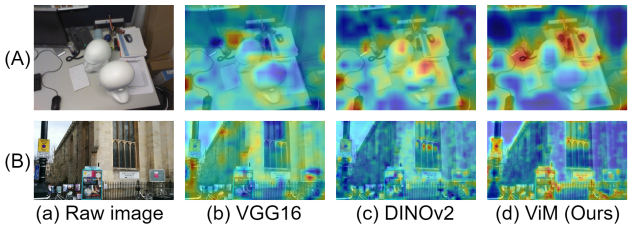

Figure 6: Attention heatmaps of CNN, ViT, and ViM.

Latency and Memory

Table 8 outlines the computational overheads of various RPE models processing a query image in 480×640 resolution. 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-MLP for rapid convergence, DeMT reduces pose regression time to just 6.09ms. In contrast, ReLoc3r incurs a significant memory overhead exceeding 100GB FLOPs due to implicit stereo computations, whereas DeMT reduces this by 2.28%.

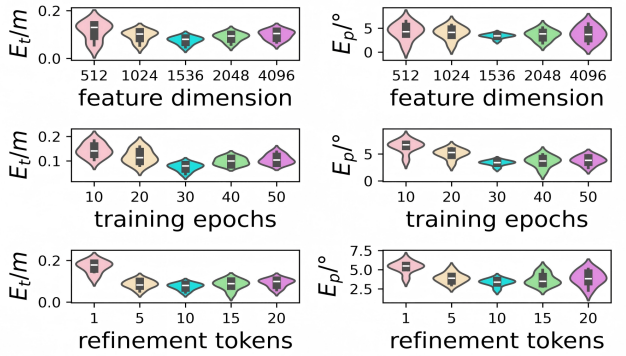

Figure 7: Parameter comparisons of DeMT: (A) feature dimension: 512, 1024, 1536, 2048, and 4096; (B) training epochs:10, 20, 30, 40, and 50; (C) reference tokens: 1, 5, 10, 15, and 20. The left column is position errors, and the right is rotation errors.

Conclusion

In this paper, we propose DeMT as an efficient solution for map-free visual relocalization. DeMT tackles RPR problem through invariant features learning and causal anchor reasoning among complex scenes. Key innovations like the CMP and SSG-MLP modules substantially reduce incorrect pose estimations. It also marks the first successful integration of domain transfer between VPR and RPR. Extensive experiments validate DeMT's precise localization accuracy, rapid speed, and robust generalization in both indoor and outdoor environments. 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), New-Generation AI Flagship Project of Guangdong Province (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.; Giguère, P.; et al. 2022. GSV-Cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513: 194–203.

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*.

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.

Izquierdo, S.; and Civera, J. 2024. Optimal transport aggregation for visual place recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17658–17668.

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.

Lu, F.; Zhang, L.; Lan, X.; et al. 2024. Towards seamless adaptation of pre-trained models for visual place recognition. In *International Conference on Learning Representations*, 1–22.

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.

Sünderhauf, N.; Neubert, P.; Protzel, P.; et al. 2013. Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons. In *IEEE International Conference on Robotics and Automation*, 2013. Citeseer.

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.

Warburg, F.; Hauberg, S.; Lopez-Antequera, M.; et al. 2020. Mapillary street-level sequences: A dataset for lifelong place recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2626–2635.

Winkelbauer, D.; Denninger, M.; Triebel, R.; et al. 2021. Learning to localize in new environments from synthetic training data. In *IEEE International Conference on Robotics and Automation*, 5840–5846.

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.