machine learning & knowledge extraction

MDPI

Article

Debiased Multiplex Tokenization Using Mamba-Based Pointers for Efficient and Versatile Map-Free Visual Relocalization †

Wenshuai Wang [1,2], Hong Liu [1,*], Shengquan Li [2], Peifeng Jiang [1], Dandan Che [2,3] and Runwei Ding [2,*]

[1] State Key Laboratory of General Artificial Intelligence, Shenzhen Graduate School, Peking University, Shenzhen 518055, China; wangws@stu.pku.edu.cn (W.W.); jpf@stu.pku.edu.cn (P.J.)
[2] Pengcheng Laboratory, Shenzhen 518055, China; lishq@pcl.ac.cn (S.L.); chedd2021@mail.sustech.edu.cn (D.C.)
[3] Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen 518055, China
* Correspondence: hongliu@pku.edu.cn (H.L.); dingrw@pcl.ac.cn (R.D.)
† This paper is an extended version of our paper published in AAAI 2026.

Abstract

Visual localization plays a critical role for mobile robots to estimate their position and orientation in GPS-denied environments. However, its efficiency, robustness, and generalization are fundamentally undermined by severe viewpoint changes and dramatic appearance variations, which present persistent challenges for image-based feature representation and pose estimation under real-world conditions. Recently, map-free visual relocalization (MFVR) has emerged as a promising paradigm for lightweight deployment and privacy isolation on edge devices, while how to learn compact and invariant image tokens without relying on structural 3D maps still remains a core problem, particularly in highly dynamic or long-term scenarios. In this paper, we propose the Debiased Multiplex Tokenizer as a novel method (termed as DMT-Loc) for efficient and versatile MFVR to address these issues. Specifically, DMT-Loc is built upon a pretrained vision Mamba encoder and integrates three key modules for relative pose regression: 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 autoregressive pointer attribution. Third, Geometry-Informed Pose Regression empowers multi-layer perceptrons with a symmetric swap gating mechanism operating inside each decoupled regression head to support accurate and flexible pose prediction in both pair-wise and multi-view modes. Extensive evaluations across seven public datasets demonstrate that DMT-Loc substantially outperforms existing baselines and ablation variants in diverse indoor and outdoor environments.

Keywords: visual localization; relative pose regression; Mamba codec; causal attention; pointer network

Check for updates

Received: 8 February 2026
Revised: 8 March 2026
Accepted: 11 March 2026
Published: 23 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

As a cornerstone of intelligent perception, visual localization (VL) aims to precisely determine the six-degree-of-freedom (6DoF) pose of a mobile robot by referencing geo-tagged imagery or preconstructed maps [1]. 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 [2]. As illustrated in Figure 1, 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. Throughout the past decade, the rich color, geometric, and semantic information inherent in RGB images has been a major driver for deep learning research in visual modeling. However, convolutional neural networks (CNNs) remain limited by their inherent receptive fields, and vision Transformers (ViTs) leverage self-attention and cross-attention to capture global context with parallel computation. Vision Mambas (ViMs) further employ state space models (SSMs) for dynamic feature filtering in linear time complexity [3]. Nonetheless, these advanced encoders can introduce inherent biases into feature representations.

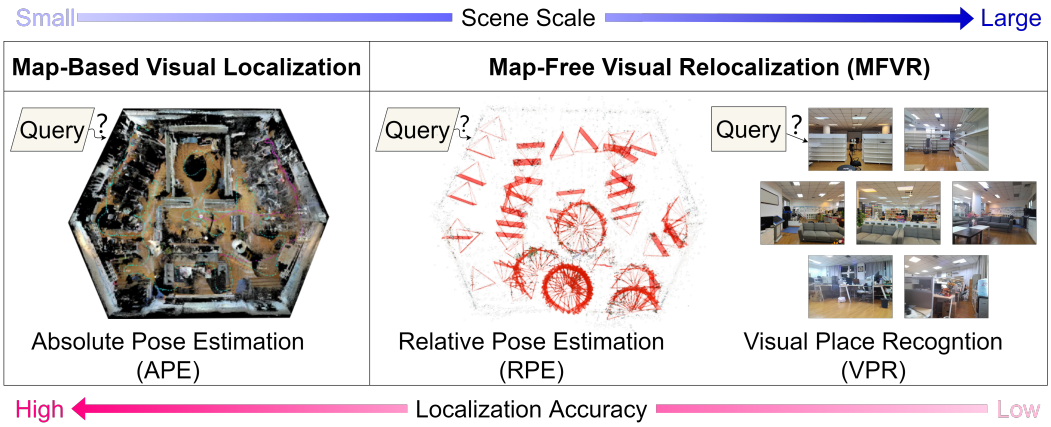


Figure 1. Main paradigms of visual localization. For a query image, absolute pose estimation (APE) matches it directly to the 3D scene map, while relative pose estimation (RPE) and visual place recognition (VPR) need to retrieve the reference dataset. Moreover, RPE is supposed to make a better trade-off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

To mitigate this issue, an effective tokenizer is essential for distilling invariant scene features into robust image tokens. This tokenization strategy integrates image retrieval and pose estimation into a unified 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. Therefore, more discriminative descriptors are expected to fuse their complementary strengths explicitly or implicitly.

It is also noteworthy that a domain gap persists between VPR's discrete outputs and RPR's continuous embeddings despite their shared reliance on image retrieval [4]. This discrepancy manifests as an out-of-distribution generalization problem between camera extrinsics, where causal attention offers a pathway to remove confounders in image tokenization [5]. The term "causal" aspect here refers to autoregressive constraints commonly used in sequence modeling (e.g., masked self-attention), rather than to statistical causal inference frameworks [6]. Incidentally, since the regressive pipeline struggles to synthesize novel viewpoints, standard multi-layer perceptrons (MLPs) lack adaptive weight modulation and geometric smoothness constraints despite their current prevalence in pose regression.

In this paper, we propose a novel causal multiplex tokenizer named DMT-Loc as an efficient MFVR framework for robust sensing and fast computing. As depicted in Figure 2, DMT-Loc integrates three core modules: (1) Multiplex Interactive Tokenization

(MxIT) enhances sparse voxel affinities in ViM-derived feature maps via refined non-local self-attention [7], synergistically addressing spatial sparsity, frequency consistency, and channel redundancy through its branch interventions for emergent robustness in feature aggregation; (2) Debiased Anchor Registration (DAR) employs hierarchical navigable small world (HNSW) graphs [8] for incremental retrieval of proximate tokens, and then creates a specific Causal Mamba Pointer (CMP) for their autoregressive attribution, ensuring the structural consistency with a visual backbone; (3) Geometry-Informed Pose Regression (GIPR) introduces a novel symmetric swap gated (SSG) MLP, which employs routing-based learning for intra-head self-sufficient feature interactions to regress 6DoF poses from single or multiple frames, demonstrating superior convergence and generalization capabilities in supervised learning.

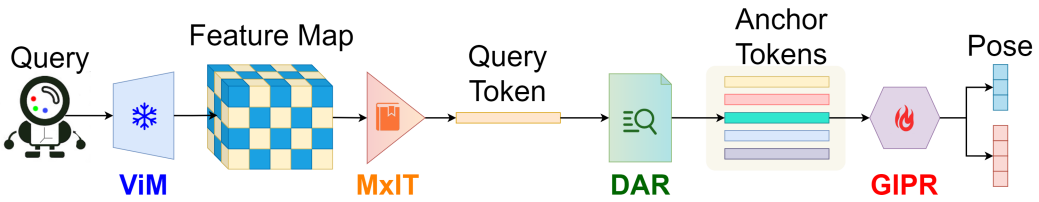


Figure 2. The pipeline of our proposed method. Four modules are designed and cascaded for relative pose regression (RPR): fronzen Pretrained ViM Encoder, Multiplex Interactive Tokenization (MxIT), Debiased Anchor Registration (DAR), and learnable Geometry-Informed Pose Regression (GIPR).

Fundamentally, DMT-Loc 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 blur or illumination changes, but also achieves linear time complexity via the coupled Mamba codec and HNSW search. Compared with a series of mainstream methods, DMT-Loc shows superiority in feature fidelity, localization accuracy, inference speed, and scene adaptability on diverse datasets, exemplified by indoor 7Scenes [9] and outdoor Cambridge Landmarks [10] datasets. In particular, all angular values are unified into unit quaternions to prevent the gimbal lock problem [11].

The main contributions of this work can be summarized as follows:

- The Debiased Multiplex Tokenizer is proposed for high-fidelity image tokenization to eliminate the mapping burden on mobile robots, leveraging multiplex interactive learning on preliminary feature maps along voxel, spatial, frequency, and channel dimensions via an enhanced non-local self-attention mechanism.
- Both tailored for token-pose debiasing that targets feature-level invariance to domain shifts, CMP efficiently scores multi-view correspondences for anchor emergence with uniform linear time complexity, while SSG-MLPs ensure decoupled and scalable pose regression aligned with geometric consistency constraints and $SE(3)$ robot kinematics.
- Comprehensive experiments on various scenes of eight VL datasets confirm the efficacy of our proposed method in both indoor and outdoor environments, especially under challenging circumstances of severe viewpoint and illumination changes.

2. Related Work

2.1. Visual Place Recognition

VPR is typically formulated as an image retrieval (IR) problem governed by similarity thresholds. NetVLAD [12] integrates a CNN with the vector of local aggregated descriptors (VLAD) [13], establishing a pioneering approach for global feature aggregation. DOLG [14] employs orthogonal projection to decompose local descriptors and explicitly concatenates them with global features. Patch-NetVLAD [15] and TransVPR [16] leverage ViTs to

construct feature pyramids for patch-level fusion. To mitigate redundancy in explicit methods, implicit techniques have gained traction. CosPlace [17] adopts generalized mean pooling (GeM) for domain adaptation, while MixVPR [18] enhances context mining through MLP-based mixer blocks. Recent approaches such as SALAD [19], SelaVPR [20], and EffoVPR [21] further fine-tune the pre-trained DINOv2 [22] features using joint adaptation of global and local features, enabling more effective assignment of static landmarks.

2.2. Absolute Pose Estimation

APE aims to establish precise stereo correspondences through map structures or neural encodings. Structure-based pipelines (SbP) [23,24] 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 [25] and PoseNet family [10], respectively. However, their effectiveness in dynamic environments is limited by mapping costs and the requirement for accurate pose regression [26]. SACNet [27], LENS [28], and Marepo [29] focus on feature compression, synthetic dataset generation, and scene generalization, respectively.

2.3. 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 [30] that heuristically solve the essential matrix, Relative Pose Regression (RPR) directly predicts pose errors between image pairs in a data-driven manner. AnchorNet [31] extends PoseNet using anchor frame clusters. NN-Net [32] examines pairwise differences via Siamese networks, while ReLocNet [33] and EssNet [34] enhance its accuracy through continual learning and geometric constraints, respectively. RelFormer [35] employs a dual-branch Transformer to address scene generalization challenges. RelPoseGNN [36] exploits multi-frame collaboration using graph neural networks (GNNs). ReLoc3r [37] builds upon the popular scene reconstructor DUSt3R [38] to incorporate multi-view geometric constraints, achieving state-of-the-art (SoTA) RPR performance on the standard datasets; however, it adapts poorly to real-world challenging environments.

3. Methodology

The overview of DMT-Loc framework is illustrated in Figure 3.

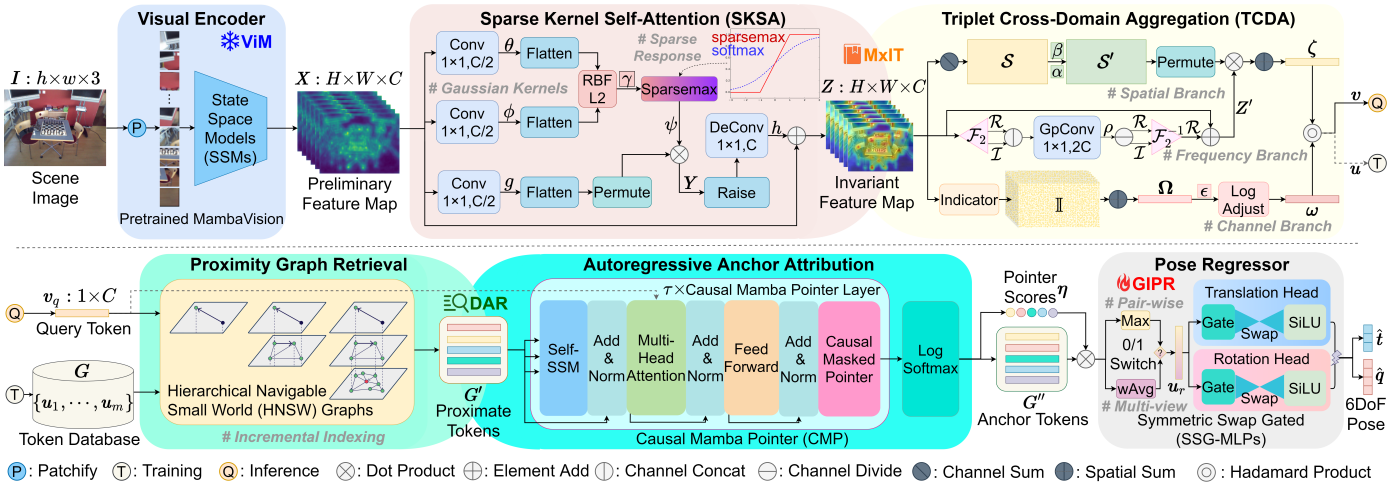


Figure 3. Overview of the DMT-Loc framework. It sequentially performs self-attentive mapping from images to tokens and causal attentive regressing to poses. RBF, DeConv, and GpConv stand for radial basis function, deconvolution, and grouped convolution, respectively.

3.1. Feature Map Encoding

Given a scene image $\boldsymbol{I} \in \mathbb{R}^{h \times w \times 3}$, a pretrained ViM model $f_{\text{vim}}(\cdot)$ is adopted 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.

Leveraging SSMs, ViM efficiently coordinates global receptive fields with linear computational complexity. A structural comparison among CNN, ViT, and ViM is provided in Figure 4.

3.2. Multiplex Interactive Tokenization

MxIT 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).

3.2.1. Sparse Kernel Self-Attention

SKSA extends vanilla self-attention via Gaussian kernels and sparse responses, incorporating a voxel-wise residual for feature enhancement.

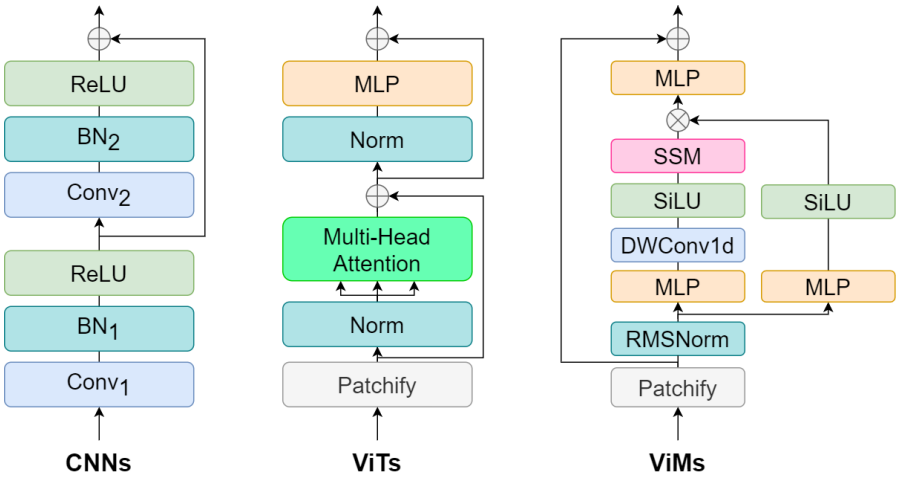


Figure 4. Structural comparison of CNN, ViT, and ViM. ViM offers the global receptive fields of ViT while maintaining linear computational complexity via state-space models (SSMs).

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

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

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

Sparse Response: To encourage the module to focus on more compact and informative features, sparsemax [39] activation is adopted to assign sparse voxel probabilities (vs. softmax) and yield zero probability for less relevant features, thereby reducing the influence of noisy affinities and leading to the SKSA response $\boldsymbol{Y} \in \mathbb{R}^{HW \times C}$ as

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

where $\psi(\cdot)$ denotes a cutoff threshold, and $g(\cdot)$ is a positional embedding function. $\boldsymbol{J}_{i,j} = \boldsymbol{1}_{i \times j}$ 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 $\boldsymbol{Z} \in \mathbb{R}^{H \times W \times C}$:

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

3.2.2. 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 $*$ refers to 2D convolution; $\mathcal{F}_2(\cdot)$ denotes 2D fast Fourier transform [40] and $\mathcal{F}_2^{-1}(\cdot)$ denotes its inverse; $\mathcal{R}(\cdot)$ and $\mathcal{I}(\cdot)$ extract the real and imaginary components, respectively; $\boldsymbol{W}_\rho$ is the weight matrix of 1×1 grouped convolution.

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

$$\boldsymbol{\Omega} = \frac{1}{HW}\boldsymbol{J}_{W,H}\mathbb{I}(\boldsymbol{Z} > \boldsymbol{0}), \tag{6}$$

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

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

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

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

3.3. Debiased Anchor Registration

DAR estimates relative pose differences between image pairs and refines their misalignments to ensure geometric consistency. This two-stage process comprises Proximity Graph Retrieval and Autoregressive Anchor Attribution.

Proximity Graph Retrieval: Given a query token $\boldsymbol{v}_q$ and a reference database $\boldsymbol{G} = \{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$ organized as proximity graphs [8], the HNSW algorithm $f_{\text{hnsw}}(\cdot, \cdot)$ is performed to retrieve the top-k candidate tokens incrementally:

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

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

This search initiates at a random entry point in the top layer of HNSW graphs and iteratively traverses to its neighbors with the smallest cosine distance to $\boldsymbol{v}_q$, descending through denser layers until k nearest neighbors in the bottom layer are identified.

Autoregressive Anchor Attribution: A causal attention mechanism is adopted to impose consistency and autoregressive constraints on Mamba decoding, so as to suppress confounding matches and distill geometrically consistent anchors.

To filter out mismatched correspondences, a τ-layer CMP $f^{\tau}_{\text{cmp}}(\cdot,\cdot)$ refines the candidate set $\boldsymbol{G'}$ into a reliable anchor set $\boldsymbol{G''}$:

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

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

CMP also outputs the normalized confidence scores $\boldsymbol{\eta} \in (0,1)^k$ for $\boldsymbol{G''}$, derived from its final hidden state after autoregressive processing.

Algorithm 1 provides the pseudocode for CMP. The self-SSM ensures its linear computational complexity, while the integrated causal masking mechanism overcomes the limited autoregressive capability of the original Mamba.

Algorithm 1 Pseudocode for Causal Mamba Pointer (CMP)

Require: Input batch proximate tokens $\boldsymbol{u}_{i'} \in \mathbb{R}^{B\times k\times C}$, batch query tokens $\boldsymbol{v}_q \in \mathbb{R}^{B\times 1\times C}$
Ensure: Output batch anchor tokens $\boldsymbol{u}_{i''} \in \mathbb{R}^{B\times k\times C}$, batch confidence scores $\boldsymbol{\eta} \in (0,1)^{B\times k}$

$\boldsymbol{x} \leftarrow \boldsymbol{u}_{i'}$
for layer $l = 1$ **to** τ **do**
  // Self-SSM block:
  $\boldsymbol{h} \leftarrow \text{LayerNorm}_1(\boldsymbol{x})$
  $\boldsymbol{h} \leftarrow \text{Mamba}(\boldsymbol{h})$ {State-space model}
  $\boldsymbol{x}_{\text{self}} \leftarrow \boldsymbol{x} + \text{Dropout}_1(\boldsymbol{h})$
  // Cross-attention:
  if $\boldsymbol{v}_q \neq$ None **then**
    $\boldsymbol{h} \leftarrow \text{LayerNorm}_2(\boldsymbol{x}_{\text{self}})$
    $\boldsymbol{m} \leftarrow \text{upper_tri}(-\infty, \text{diagonal} = 1)$ {Causal mask}
    $\boldsymbol{h} \leftarrow \text{MultiHeadAttn}(\boldsymbol{h}, \boldsymbol{v}_q, \boldsymbol{v}_q, \boldsymbol{m})$ {Autoregressive refinement}
    $\boldsymbol{x}_{\text{cross}} \leftarrow \boldsymbol{x}_{\text{self}} + \text{Dropout}_2(\boldsymbol{h})$
  end if
  // Feed-forward network:
  $\boldsymbol{h} \leftarrow \text{LayerNorm}_3(\boldsymbol{x}_{\text{cross}})$
  $\boldsymbol{h} \leftarrow \text{FFN}(\boldsymbol{h})$ {Linear $\rightarrow$ Activation $\rightarrow$ Linear}
  $\boldsymbol{x} \leftarrow \boldsymbol{x}_{\text{cross}} + \text{Dropout}_3(\boldsymbol{h})$
  // Last layer:
  if layer $l == \tau$ **then**
    $\boldsymbol{\eta} \leftarrow \text{softmax}(\text{Linear}_{C\rightarrow 1}(\boldsymbol{x}))$ {Confidence scores}
  end if
end for
$\boldsymbol{u}_{i''} \leftarrow \text{LayerNorm}_4(\boldsymbol{x})$
return $\boldsymbol{u}_{i''}, \boldsymbol{\eta}$

3.4. Geometry-Informed Pose Regression

Given an anchor token $\boldsymbol{u}_r$, 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 detailed in Algorithm 2, each regression head performs an inner SSG operation on $\boldsymbol{u}_r$, which is first split into gate and value partitions, then modulated bidirectionally, fused via learnable routing weights, and finally projected to the output.

Therefore, the predicted absolute pose $\hat{\boldsymbol{p}}_q^{(r)} = [\hat{\boldsymbol{t}}_q^{(r)}, \hat{\boldsymbol{q}}_q^{(r)}]$ of the query image $\boldsymbol{I}_q$ can be computed using the known pose $\boldsymbol{p}_r = [\boldsymbol{t}_r, \boldsymbol{q}_r]$ of $\boldsymbol{u}_r$ as

$$\hat{\boldsymbol{t}}_q^{(r)} = \boldsymbol{t}_r + \Delta\boldsymbol{t}, \ \hat{\boldsymbol{q}}_q^{(r)} = \boldsymbol{q}_r \otimes \Delta\boldsymbol{q}, \tag{10}$$

where $\otimes$ denotes quaternion multiplication.

Algorithm 2 Pseudocode for symmetric swap gated multi-layer perceptrons (SSG-MLPs)

Require: Input anchor token $\boldsymbol{u}_r \in \mathbb{R}^C$
Ensure: Output relative pose difference $\Delta \boldsymbol{p} \in \mathbb{R}^7$ (3 translation + 4 quaternion)
// SSG-MLP 1 for translation regression:
$\boldsymbol{g}_t, \boldsymbol{v}_t \leftarrow \text{Split}(\text{Linear}_{C \to 2h}(\boldsymbol{u}_r))$ {Split into gate and value partitions}
$\boldsymbol{p}_t^1 \leftarrow \text{SiLU}(\boldsymbol{g}_t) \odot \boldsymbol{v}_t$ {Gate to value modulation}
$\boldsymbol{p}_t^2 \leftarrow \text{SiLU}(\boldsymbol{v}_t) \odot \boldsymbol{g}_t$ {Symmetric swap}
$\boldsymbol{w}_t \leftarrow \text{softmax}(\boldsymbol{\Theta}_t)$ {Learnable routing weights}
$\boldsymbol{z}_t \leftarrow w_{t1}\boldsymbol{p}_t^1 + w_{t2}\boldsymbol{p}_t^2$
$\Delta \boldsymbol{t} \leftarrow \text{Linear}_{h \to 3}(\boldsymbol{z}_t)$
// SSG-MLP 2 for rotation regression:
$\boldsymbol{g}_q, \boldsymbol{v}_q \leftarrow \text{Split}(\text{Linear}_{C \to 2h}(\boldsymbol{u}_r))$ {Split into gate and value partitions}
$\boldsymbol{p}_q^1 \leftarrow \text{SiLU}(\boldsymbol{g}_q) \odot \boldsymbol{v}_q$ {Gate to value modulation}
$\boldsymbol{p}_q^2 \leftarrow \text{SiLU}(\boldsymbol{v}_q) \odot \boldsymbol{g}_q$ {Symmetric swap}
$\boldsymbol{w}_q \leftarrow \text{softmax}(\boldsymbol{\Theta}_q)$ {Learnable routing weights}
$\boldsymbol{z}_q \leftarrow w_{q1}\boldsymbol{p}_q^1 + w_{q2}\boldsymbol{p}_q^2$
$\Delta \boldsymbol{q} \leftarrow \text{Linear}_{h \to 4}(\boldsymbol{z}_q)$
// Concatenate outputs for 6DoF pose:
$\Delta \boldsymbol{p} \leftarrow [\Delta \boldsymbol{t},\ \Delta \boldsymbol{q}]$
return $\Delta \boldsymbol{p}$

3.5. Loss and Inference

3.5.1. Training Loss

The RPR loss L_{rpr} is defined as the L_1 norm of pose differences:

$$L_{\text{rpr}} = ||\boldsymbol{t}_q - \hat{\boldsymbol{t}}_q||_1 + ||2 \arccos |\langle \boldsymbol{q}_q, \hat{\boldsymbol{q}}_q \rangle| \,||_1. \tag{11}$$

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

Overall, the total loss L_{total} is a weighted sum of three terms:

$$L_{\text{total}} = L_{\text{rpr}} + \lambda L_{\text{hnsw}} + \mu L_{\text{cmp}}, \tag{12}$$

where $\lambda, \mu > 0$ are balance coefficients, and $L_{\text{hnsw}}, L_{\text{cmp}}$ are the negative log-likelihood (NLL) losses that encourage the true reference tokens for $\boldsymbol{v}_q$ to be ranked at the top positions of $\boldsymbol{G}'$ and $\boldsymbol{G}''$, respectively.

3.5.2. Online Inference

DMT-Loc supports for two RPR modes of online MFVR inference: pair-wise calibration and multi-view association.

Pair-wise Mode: Only the top-ranked anchor $\boldsymbol{u}_{1''}$ is selected from $\boldsymbol{G}''$, so the absolute pose error is computed as

$$E_{pw}(\boldsymbol{v}_q, \boldsymbol{u}_{1''}) = (||\boldsymbol{t} - \hat{\boldsymbol{t}}||_2,\ 2 \arccos |\boldsymbol{q}\hat{\boldsymbol{q}}| \frac{180^\circ}{\pi}). \tag{13}$$

Multi-view Mode: Leveraging the confidence scores $\boldsymbol{\eta} \in (0,1)^k$ produced by CMP, the final pose error is obtained as a weighted average over the top-k anchors:

$$E_{mv}^k(\boldsymbol{v}_q, \boldsymbol{G''}) = \sum_{i=1}^{k} \boldsymbol{\eta}_i E_{pw}(\boldsymbol{v}_q, \boldsymbol{u}_{i''}). \tag{14}$$

4. Experiments

4.1. Datasets

To validate the effectiveness of the proposed DMT-Loc framework and its components, we conduct comprehensive experiments on a variety of public benchmark datasets, including four APE/RPE datasets (i.e., 7Scenes [9], Cambridge Landmarks [10], InLoc [41], and Aachen Day-Night v1.1 [42]) and three VPR datasets (i.e., Pittsburgh-250k [12], MSLS [43], and Nordland [44]). As detailed in Table 1, APE/RPE datasets encompass both indoor and outdoor environments and vary significantly in scale and conditions, while VPR datasets are mainly captured from large-scale streetscapes. All datasets are publicly available to ensure experimental reproducibility.

Table 1. Details of the adopted VL datasets. "★" and "☆" indicate presence and absence, respectively. Note that the MSLS dataset provides two distinct splits (i.e., MSLS-val and MSLS-challenge).

	Dataset	# Refer.	# Query	Motion	Light	Season	Occlusion
APE/RPE	7Scenes [9]	26.0k	17.0k	★★☆	★☆☆	☆☆☆	☆☆☆
	Cambridge [10]	8.4k	4.8k	★★★	★★☆	☆☆☆	★★☆
	InLoc [41]	10.0k	329	★★☆	★★☆	☆☆☆	★★★
	Aachenv1.1 [42]	6.7k	1.0k	★★☆	★★★	☆☆☆	★★★
VPR	Pitts250k-test [12]	83.9k	8.2k	★★☆	★★☆	★☆☆	★★☆
	MSLS-val [43]	18.9k	740	★★★	★☆☆	★★☆	★☆☆
	MSLS-challenge [43]	38.8k	27.1k	★★☆	★★★	★★★	★★★
	Nordland-test [44]	27.6k	3.5k	☆☆☆	★☆☆	★★★	☆☆☆

4.2. 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 medians, averages, standard deviations, and accuracies under specific thresholds. VPR performance is quantified by Recall@1/5/10, calculated as the ratio of correct matches among the top-ranked reference images.

4.3. Implementation Details

4.3.1. Models

MambaVision-L [45] pretrained on the ImageNet21K dataset [46] is chosen as the visual encoder with fixed feature channel dimension $C = 1536$, removing its final average pooling and fully connected layers to extract preliminary feature maps for scene images.

4.3.2. Hyperparameters

For MxIT, $\theta(\cdot)$ and $\phi(\cdot)$ in Equation (1) are implemented by 1×1 radial basis function (RBF) convolution [47], while $g(\cdot)$ is implemented by 1×1 bottleneck convolution. α and β in Equation (4) are set to 0.5 and 2, respectively. ϵ in Equation (7) is set to 1×10^{-4}. 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. In addition, the loss balance coefficients λ and μ in Equation (12) are both set to 1.

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.

4.3.3. 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 1.11.0 with Python 3.8 and CUDA 12.1.

4.4. Quantitative Comparisons

4.4.1. RPR Performance on Benchmark Datasets

Table 2 presents a comprehensive evaluation of DMT-Loc on RPR performance, alongside representative VL methods on the 7Scenes and Cambridge Landmarks datasets. For pairwise calibration, DMT-Loc achieves SoTA results among pair-wise RPR methods, with average errors of 7.3 cm and 3.25° for indoor scenes and 15.5 cm and 1.02° for outdoor scenes. These results underscore DMT-Loc'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 Marepo, DMT-Loc 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.

Table 2. Median errors (cm/ °) of baseline methods on benchmark VL datasets. "↓" indicates that lower values are better. Best RPR results are highlighted in bold and second best are 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 to our effort.

		7Scenes (Indoor, Normal) ↓								Cambridge Landmarks (Outdoor, Normal) ↓					
	Method	**Chess**	**Fire**	**Heads**	**Office**	**Pumpkin**	**Kitchen**	**Stairs**	**Average**	**College**	**Hospital**	**Shop**	**Church**	**Average-4**	**Court**
SbP	AS [23]	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 [48]	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 [24]	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 [49]	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-star [25]	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 [26]	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 [50]	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++[51]	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 [52]	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 † [27]	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 [10]	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 [53]	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 [54]	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 † [28]	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 † [55]	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	–/–	–/–	–/–	–/–	–/–	–/–
	Marepo [29]	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 [32]	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 [33]	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 [31]	**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 [34]	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 [56]	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 [35]	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
	DMT-Loc (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 † [57]	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 [36]	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 [37]	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
	DMT-Loc (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**

4.4.2. MFVR Performance on Challenging Datasets

Several competitive methods are further evaluated on more challenging datasets to assess their MFVR benefits, as detailed in Table 3. On the difficult indoor InLoc dataset, DMT-Loc attains accuracies of 64.9%/84.8%/91.3% on DUC1 and 75.6%/88.3%/92.8% on DUC2 under the (0.25/0.5/1.0 m, 10°) threshold, surpassing SoTA performance among all competitive methods. For challenging outdoor VL, map-based methods such as HLoc and DeViLoc achieve excellent accuracy on the Aachenv1.1 dataset. However, DMT-Loc still delivers highly competitive performance in both daytime and nighttime conditions with significantly lower computational overhead. Specifically, it reduces storage requirements by two orders of magnitude (0.2 GB vs. 15.9 GB for DeViLoc) and cuts training time to

under 2.5 h compared to over 99 h for most map-based competitors. These results highlight DMT-Loc's effectiveness and universality in complex workspaces.

Table 3. Average accuracies (%) and training overheads of competitive methods on challenging VL datasets. "↑" indicates that higher values are better, and vice versa for "↓". Overall best/second best results are marked in blue/green. Best/second best overheads are highlighted in bold/underlined.

	Method	InLoc (Indoor, Difficult) Acc.@(0.25/0.5/1.0 m, 10°) ↑ DUC1	InLoc (Indoor, Difficult) Acc.@(0.25/0.5/1.0 m, 10°) ↑ DUC2	Aachenv1.1 (Outdoor, Difficult) Acc.@(0.25/0.5/5.0 m, 2/5/10°) ↑ Day	Aachenv1.1 (Outdoor, Difficult) Acc.@(0.25/0.5/5.0 m, 2/5/10°) ↑ Night	Training Time ↓ (Hours)	Storage Size ↓ (GB)
VPR	VLAD [13]	0.20/12.5/18.7	0.30/13.8/19.1	0.00/0.10/22.8	0.00/1.00/19.4	<u>18</u>	5.2
	NetVLAD [12]	8.20/24.7/48.9	6.40/26.3/54.9	0.00/0.20/18.9	0.00/0.00/14.3	30	4.8
	VLAD+In. [58]	6.40/26.3/50.9	10.3/32.3/61.5	0.00/0.20/22.1	0.00/1.00/22.4	34	7.1
APE	AS [23]	–/–/–	–/–/–	85.3/92.2/97.9	39.8/49.0/64.3	60	<u>3.2</u>
	D2Net [59]	44.4/58.6/71.2	31.3/49.6/67.9	84.8/92.6/97.5	84.7/90.8/96.9	≥72	22.3
	PixLoc [48]	25.5/47.3/68.8	32.4/54.7/79.5	74.3/79.3/87.4	61.0/65.8/79.3	≥72	48.5
	HSCNet++ [51]	–/–/–	–/–/–	72.7/81.6/91.4	43.9/57.1/76.5	≥72	27.4
	DeViLoc [24]	55.8/63.8/88.9	61.0/72.5/89.4	87.4/94.8/98.2	87.8/93.9/100.0	≥72	15.9
RPE	HLoc (SP+SG) [30]	49.0/68.7/80.8	53.4/77.1/82.4	89.6/95.4/98.8	86.7/93.9/100.0	–	35.7
	RelFormer [35]	51.5/73.7/86.4	55.0/74.0/81.7	60.2/67.1/78.5	51.4/62.5/73.5	27	8.5
	ReLoc3r [37]	59.6/79.3/90.9	71.2/87.0/91.6	61.5/77.0/89.6	53.8/63.7/75.8	85	4.9
	DMT-Loc (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	**≤2.5**	**0.2**

4.4.3. VPR Performance

As listed in Table 4, DMT-Loc is further compared across a series of VPR benchmarks, which consistently outperforms existing approaches on both large-scale urban datasets (Pitts250k-test, MSLS-val) and challenging diurnal or seasonal environments (MSLS-challenge, Nordland-test). These results confirm that the image tokenization scheme provides a more discriminative and generalizable representation than global–local fusion techniques commonly used in VPR.

Table 4. Recall@1/5/10 (%) comparisons on VPR datasets. "↑" indicates that higher values are better. Best results are in bold and second best are underlined.

Method	Pitts250k-Test ↑	MSLS-Val ↑	MSLS-Challenge ↑	Nordland-Test ↑
NetVLAD [12]	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 [14]	89.9/95.4/96.7	82.0/88.9/91.4	<u>75.6</u>/<u>87.1</u>/<u>90.8</u>	51.3/66.8/69.8
CosPlace [17]	88.4/94.5/95.7	82.8/89.7/92.0	61.4/72.0/76.6	54.4/69.8/75.9
Patch-NetVLAD [15]	87.5/94.5/96.0	79.5/86.2/87.7	48.1/57.6/60.5	44.9/50.2/52.2
TransVPR [16]	89.0/94.9/96.2	86.8/91.2/92.4	63.9/74.0/77.5	<u>61.3</u>/71.7/75.6
MixVPR [18]	91.5/95.5/96.3	88.0/92.7/94.6	64.0/75.9/80.6	58.4/<u>74.6</u>/<u>80.0</u>
SelaVPR [20]	<u>92.7</u>/<u>98.0</u>/<u>98.9</u>	<u>87.7</u>/<u>95.8</u>/<u>96.6</u>	69.6/86.9/90.1	47.2/66.6/74.1
DMT-Loc (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**

4.5. Qualitative Visualization

4.5.1. Instances of Visual Localization

Figure 5 illustrates the visualization results of DMT-Loc across five query images, representing VL challenges such as viewpoint ambiguity, motion blur, illumination shifts, and dynamic occlusions, respectively. DMT-Loc 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 space of each scene to calibrate these camera frustums and visualize their pose errors as shown in column (g).

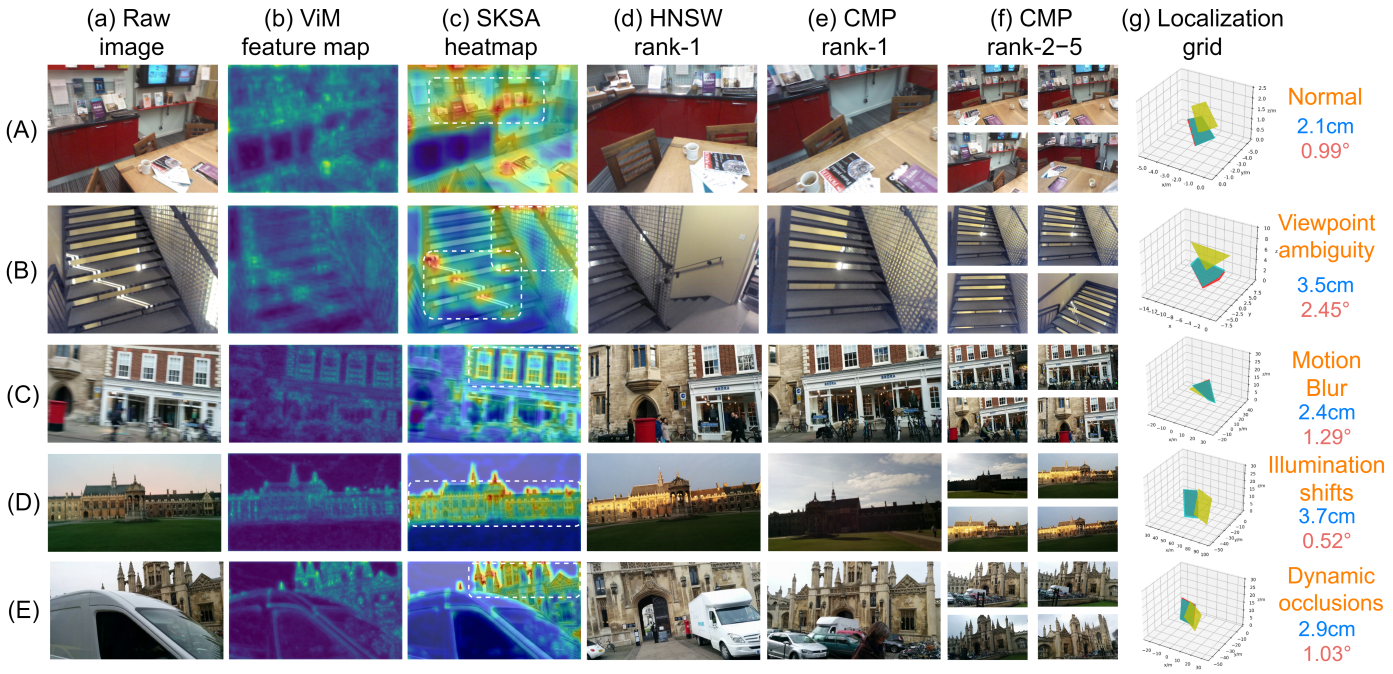


Figure 5. An illustration of DMT-Loc's pipeline on representative queries from 7Scenes dataset (rows **A**,**B**) and Cambridge Landmarks dataset (rows **C**–**E**). From left to right: (**a**) raw image; (**b**) feature map from the ViM encoder; (**c**) attention heatmap from the SKSA module, with key regions highlighted by white dashed boxes; (**d**–**f**) retrieval results showing the top-1 match from HNSW (**d**), the re-ranked top-1 (**e**), and top-2–5 (**f**) matches from the CMP module; (**g**) 3D localization grid, with camera frustums indicating ground truth (red), nearest reference (yellow), and predicted pose (cyan). The main challenges and localization errors are indicated at the far right.

In addition, Figure 6 compares the top-1 retrieval results for challenging queries with severe appearance variations. DMT-Loc correctly retrieves the corresponding reference images under both day–night (MSLS-challenge) and winter–fall (Nordland-test) transformations, while SelaVPR and NetVLAD fail in both scenarios. This demonstrates DMT-Loc's superior robustness to illumination and seasonal changes, challenges where existing global description and transformer-based methods often fail.

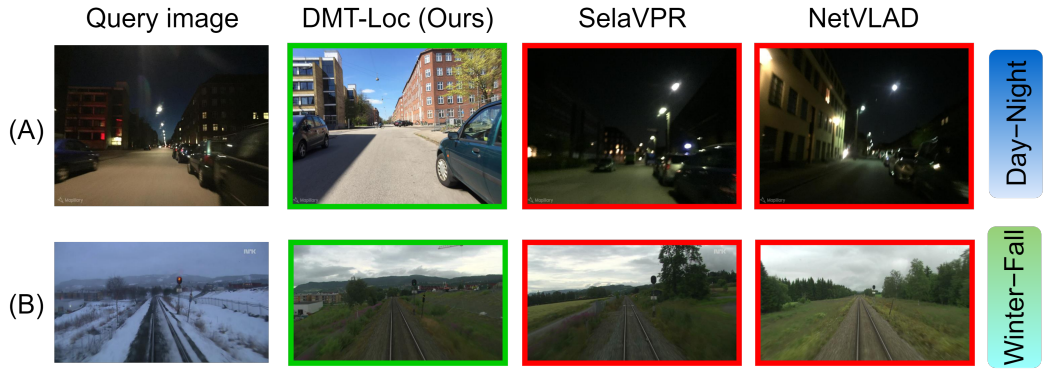


Figure 6. Rank-1 examples of VPR methods under challenging condition changes. (**A**) Query image from MSLS-challenge dataset with day–night transformation. (**B**) Query image from Nordland-test dataset with winter–fall transformation. Each method returns its most similar database image, with correct matches framed in green and incorrect matches in red.

4.5.2. Statistics of Image Tokenization

The storage efficiency of DMT-Loc is analyzed with a fixed feature dimension of 1536. As shown in Figure 7, the image tokenization scheme achieves significant compression

ratios across eight VL datasets. For instance, the 7Scenes dataset (23.6 GB) is compressed to 287.4 MB of tokens and 163.6 MB of HNSW graphs, while the Cambridge Landmarks dataset (38.5 GB) is reduced to 88.2 MB of tokens and 52.7 MB of HNSW graphs. Similar compression rates are observed for other datasets, demonstrating that DMT-Loc enables lightweight deployment for MFVR while preserving essential geometric information.

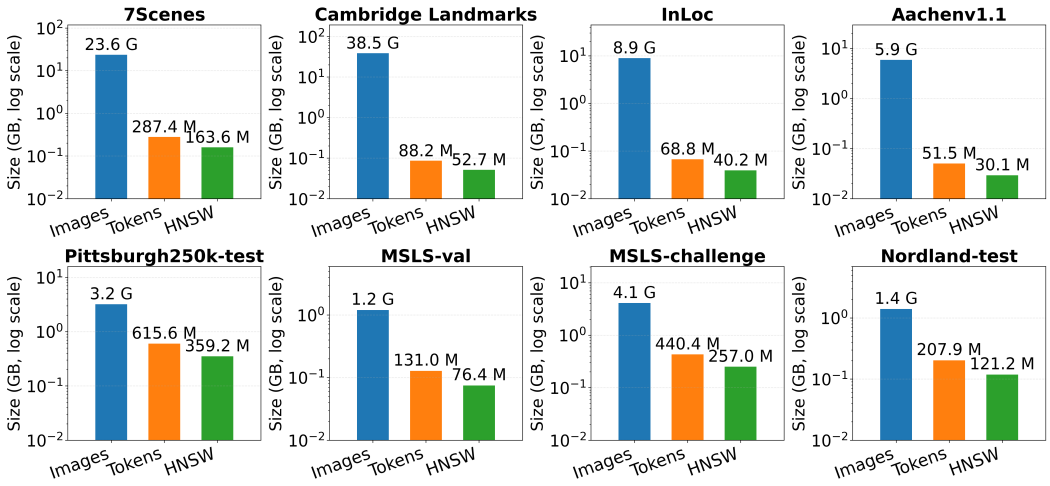


Figure 7. Storage efficiency of image tokenization across VL benchmarks with feature dimension 1536. The logarithmic y-axis scale intuitively reflects that the proposed method compresses raw datasets by orders of magnitude into compact tokens and HNSW graphs for efficient MFVR.

4.5.3. Calibrations of Camera Frustums

Based on dense voxel reconstruction of each scene of 7Scenes dataset, we calibrate the corresponding camera poses within the reconstructed map as shown in Figure 8. This joint visualization not only illustrates the accuracy of pose estimation but also provides an intuitive assessment of geometric consistency across the entire test sequence, thereby supporting the robustness of DMT-Loc for MFVR.



Figure 8. Qualitative results of camera calibration by DMT-Loc on the 7Scenes dataset. Each voxel map is reconstructed with 1000 frames per scene. Ground-truth and predicted 6DoF poses are visualized as red and green camera frustums, respectively.

5. Discussions

5.1. Efficiency Analysis

5.1.1. Efficiency of Image Retrieval

To evaluate the efficiency of DMT-Loc for fast and incremental IR, HNSW is compared against brute-force k-nearest neighbor (kNN) and other two popular approximate nearest

neighbor algorithms, i.e., k-dimension tree (K-D Tree) [60] and locality-sensitive hashing (LSH) [61]. As shown in Figure 9, the recall–latency curves on the MSLS-challenge dataset demonstrate the superior performance of HNSW search. At 5000 queries, HNSW achieves a 520× speedup over brute-force kNN (0.43 s vs. 275.43 s) while maintaining a high recall of 0.91, compared to 0.93 for kNN. In contrast, K-D Tree and LSH are significantly faster than kNN (62× and 237×, respectively) but suffer from lower recall rates (0.80 and 0.72). These results highlight HNSW's optimal balance between speed and accuracy for large-scale IR, making it well-suited for real-time VPR tasks.

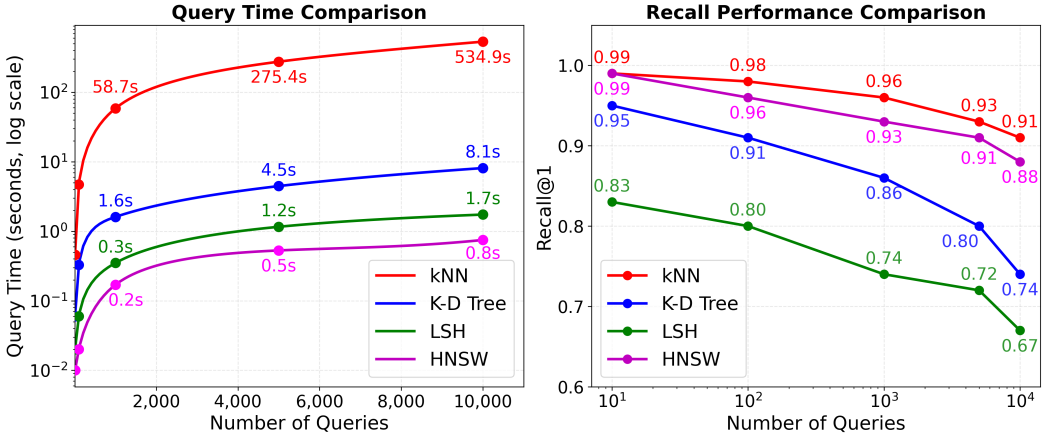


Figure 9. Performance comparison of IR algorithms on MSLS-challenge dataset. HNSW achieves the best trade-off between speed and accuracy.

5.1.2. Efficiency of Visual Encoders

Figure 10 visualizes attention heatmaps of VGG16, DINOv2, and our ViM encoder for indoor and outdoor scenes. ViM focuses on distinctive objects (e.g., telephones, billboards) while ignoring repetitive patterns (e.g., sculpture heads) and textureless regions (e.g., building walls). In contrast, VGG16 and DINOv2 attend to less informative areas. This explains the superior performance of DMT-Loc (using ViM) as listed in Tables 2 and 4.

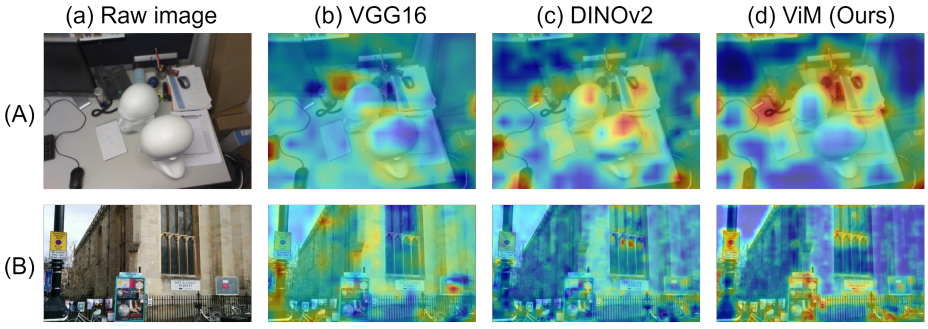


Figure 10. Attention heatmaps of visual encoders on indoor (**A**) and outdoor (**B**) scenes. (**a**–**d**) correspond to the raw images and heatmaps of VGG16, DINOv2, and ViM (ours), respectively. Warmer (cooler) colors correspond to higher (lower) attention.

5.2. Generality Analysis

5.2.1. Descriptor Transfer

To investigate the transferability of VPR descriptors to the RPR task, a set of VPR descriptors are integrated with the GIPR module to evaluate their 6DoF poses. As shown in Table 5, the prominent advantage of DMT-Loc is consistent across 7Scenes and Cambridge Landmarks datasets. Notably, global descriptors from retrieval-oriented models such as NetVLAD or MixVPR exhibit a clear geometric inconsistency when applied to continuous pose regression, whereas the MxIT module of DMT-Loc preserves structural cues with

feature reusing that are essential for both VPR and RPR. In addition, it is proven that the DAR module could provide more geometrically informative and transferable reference tokens for efficient MFVR.

Table 5. Median errors (cm/°) of different descriptors (transfered with GIPR) on 7Scenes and Cambridge Landmarks datasets. "↓" indicates that lower values are better. Best results are highlighted in bold and second best are underlined.

	7Scenes (Indoor, Normal) ↓							
Method	**Chess**	**Fire**	**Heads**	**Office**	**Pumpkin**	**Kitchen**	**Stairs**	**Average**
NetVLAD [12]	17/8.47	19/10.3	14/11.2	23/10.1	24/10.8	21/10.0	18/6.27	19.4/9.58
DOLG [14]	43/11.5	43/13.2	25/14.6	40/17.5	35/11.1	54/15.0	30/7.79	38.6/12.9
CosPlace [17]	28/14.3	23/11.2	17/14.9	27/11.1	25/13.4	26/10.9	24/9.44	24.3/12.2
Patch-NV [15]	21/10.9	29/11.5	15/11.7	29/12.2	25/9.19	24/13.1	18/7.11	23.0/10.8
MixVPR [18]	24/14.3	17/10.6	16/12.4	33/16.9	23/10.7	27/9.64	19/7.03	22.7/11.6
SelaVPR [20]	26/15.1	22/16.8	15/14.7	32/19.5	21/13.6	37/21.4	16/8.90	24.1/15.7
DMT-Loc (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**
	Cambridge Landmarks (Outdoor, Normal) ↓							
Method	**College**	**Hospital**	**Shop**	**Church**	**Average-4**	**Court**	**Average-5**	**Street**
NetVLAD [12]	182/5.40	109/5.12	66/8.28	150/10.3	127/7.28	621/15.9	226/9.00	–/–
DOLG [14]	392/6.73	218/6.01	291/10.4	416/10.7	329/8.46	969/17.3	457/10.2	–/–
CosPlace [17]	96/4.50	119/6.05	71/6.77	111/6.72	99/6.01	157/4.92	111/5.79	–/–
Patch-NV [15]	228/5.48	119/5.96	139/8.91	237/12.2	181/8.14	840/16.2	313/9.74	955/24.8
MixVPR [18]	61/4.52	46/4.27	32/4.77	48/9.43	47/5.75	56/5.88	49/5.77	356/24.3
SelaVPR [20]	59/5.95	48/5.09	37/5.83	51/7.52	49/6.10	64/6.17	52/6.11	–/–
DMT-Loc (Ours)	**23/1.19**	**18/1.02**	**9/0.87**	**12/0.98**	**15.5/1.02**	**24/0.94**	**17/1.00**	**33/2.68**

5.2.2. Scene Generalization

Furthermore, the generalization capability of DMT-Loc is evaluated under a leave-one-scene-out training protocol. As shown in Table 6, DMT-Loc substantially outperforms all prior RPR baselines on both benchmark datasets under the more challenging scene-agnostic settings, where ReLoc3r is not included since it relies on DUSt3R's inherent pre-training on large-scale multi-scene corpora. It is worth noting that the multi-view mode exhibits a slight performance degradation on unseen scenes compared to its pair-wise counterpart on the 7Scenes dataset, which can be attributed to the increased likelihood of false-positive matches when aggregating multiple anchors in an unfamiliar environment. Nevertheless, DMT-Loc-mv still achieves competitive performance (i.e., 0.59 m/0.72°) on Cambridge Landmarks dataset, underscoring the robustness of our design.

Table 6. Average errors of scene-specific/agnostic RPR on 7Scenes and Cambridge Landmarks datasets. "↓" indicates that lower values are better. Best results are highlighted in bold and second best are underlined. "–" means the value is unconverged or unobtainable.

	7Scenes (m/°) ↓		Cambridge (m/°) ↓	
Method	**Scene-Specific**	**Scene-Agnostic**	**Scene-Specific**	**Scene-Agnostic**
NN-Net [32]	0.21/9.30	0.36/18.4	–/–	–/–
ReLocNet [33]	0.21/6.73	0.29/11.3	–/–	–/–
EssNet [34]	0.22/8.03	0.89/40.2	1.08/3.42	10.4/85.8
NC-EssNet [34]	0.21/7.50	0.82/26.2	0.85/2.83	7.98/24.4
RelPoseGNN [36]	0.16/5.24	0.36/13.6	1.68/3.60	–/–
Relformer [35]	0.18/6.27	0.30/8.53	1.37/2.30	3.35/10.7
DMT-Loc-pw (Ours)	0.07/3.25	**0.16/6.09**	0.17/1.00	1.59/4.36
DMT-Loc-mv (Ours)	**0.03/1.12**	0.22/7.65	**0.08/ 0.46**	**0.59/0.72**

5.3. Ablation Study

5.3.1. Ablations on Model Components

Four key components (i.e., SKSA, TCDA, CMP, and SSG) are eliminated in turn to evaluate their contributions to the proposed method. As shown in Table 7, removing any component leads to noticeable performance degradation. The absence of CMP results in the largest error increase, particularly for orientation estimation (e.g., from 1.00° to 3.47°

on Cambridge), highlighting its critical role in fusing complementary cues and reducing ambiguities between similar poses. Disabling SSG also leads to consistent performance drop (e.g., translation errors increase by ∼30% and rotation errors by ∼75% on 7Scenes), confirming its effectiveness in geometric feature interaction for pose regression. Removing SKSA or TCDA similarly raises translation and rotation errors, demonstrating that both the symmetric attention mechanism and token-level feature aggregation contribute to robust feature representation. The entire model, integrating all four components, consistently achieves the lowest errors across both indoor and outdoor datasets, showing their synergistic enhancement of relative pose estimation.

Table 7. Average errors ± standard deviations of DMT-Loc's model components for ablation study on 7Scenes and Cambridge Landmarks datasets. "↓" indicates that lower values are better. Best results are highlighted in bold and second best are underlined. "✓" denotes module presence.

SKSA	TCDA	CMP	SSG	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
✓	✓	✓		<u>0.095</u> ± <u>0.028</u>/ <u>4.18</u> ± <u>0.73</u>	<u>0.183</u> ± <u>0.079</u>/<u>1.75</u> ± <u>0.39</u>

5.3.2. Ablations on Design Strategies

Table 8 validates the core design of DMT-Loc. Under the same ImageNet-21K pretraining, the ViM encoder outperforms CNN (VGG16, ResNet50) and ViT (DINOv2) backbones, confirming its superiority for feature extraction. The notably higher errors of DUSt3R underscore that geometric reconstruction priors alone are insufficient for robust feature aggregation. Within MxIT, the combined SKSA and TCDA design is proven to be essential; either ablating or replacing them with standard pooling (VLAD, GeM) significantly degrades performance. For DAR, HNSW offers an optimal efficiency–accuracy balance, while CMP emerges as crucial for robust matching compared to standard decoder alternatives. In GIPR, accurate pose estimation hinges on the SSG mechanism, with its symmetric swap and gating operations being more critical than specific activation function. Overall, these results demonstrate that each component uniquely contributes to DMT-Loc's superior performance, with the full framework achieving the lowest errors and highest stability across both indoor and outdoor benchmarks.

5.3.3. Ablations on Key Hyperparameters

Figure 11 presents a systematic analysis of key hyperparameters in DMT-Loc. The results indicate that optimal performance is achieved with a feature dimension of 1536, 30 training epochs, 10 reference tokens for refinement, and a dual-path swap configuration. This combination effectively balances representational capacity, training efficiency, and geometric interaction. Notably, the SSG-MLP with two swap paths outperforms both single- and multi-path variants, enabling complementary feature reasoning without introducing redundancy or excessive computation. These well-tuned hyperparameters collectively contribute to DMT-Loc's strong performance in terms of accuracy, speed, and generalization across diverse environments.

Table 9 further quantifies the sensitivity of several modular hyperparameters, reporting the maximum relative change in median pose error on the 7Scenes and Cambridge Landmarks datasets when varying each hyperparameter across its tested range while keeping others at default values. The results demonstrate consistently low sensitivity for each hyperparameter, while all of them are kept constant across datasets without per-dataset tun-

ing, confirming that DMT-Loc's performance stems from architectural design rather than overfitting. This aligns with its strong generalization in leave-one-scene-out evaluation.

Table 8. Average errors ± standard deviations of DMT-Loc's design strategies for ablation study on 7Scenes and Cambridge Landmarks datasets. "↓" indicates that lower values are better. Best results are highlighted in bold and second best are underlined.

Module		Technique	7Scenes (m/°) ↓	Cambridge (m/°) ↓
Encoder	ViM	VGG16	0.129 ± 0.026/5.96 ± 0.89	0.396 ± 0.102/3.12 ± 0.77
		ResNet50	0.119 ± 0.023/5.42 ± 0.81	0.331 ± 0.094/2.75 ± 0.52
		DINOv2-ViT	0.101 ± 0.022/4.66 ± 0.75	0.273 ± 0.083/2.45 ± 0.23
		DUSt3R	0.096 ± 0.020/4.57 ± 0.72	0.315 ± 0.086/2.59 ± 0.44
MxIT	SKSA	NLSA	0.103 ± 0.028/4.74 ± 0.71	0.212 ± 0.085/1.82 ± 0.60
		w/o RBFConv	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 ± 0.61	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
		w/o Spatial	0.142 ± 0.026/6.62 ± 1.03	0.251 ± 0.108/2.58 ± 0.82
		w/o Frequency	0.096 ± 0.041/4.59 ± 1.15	0.207 ± 0.086/1.74 ± 0.93
		w/o Channel	0.129 ± 0.025/5.13 ± 0.77	0.235 ± 0.097/2.05 ± 0.59
DAR	HNSW	kNN	<u>0.076</u> ± <u>0.018</u>/**3.14** ± <u>0.58</u>	<u>0.174</u> ± **0.057**/<u>1.21</u> ± <u>0.16</u>
		K-D Tree	0.083 ± 0.024/3.82 ± 0.65	0.192 ± 0.074/1.48 ± 0.23
		LSH	0.095 ± 0.029/4.37 ± 0.78	0.236 ± 0.092/1.86 ± 0.31
	CMP	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 ± 0.61	0.194 ± 0.083/1.33 ± 0.26
GIPR	SSG	*w/o* Gate	0.121 ± 0.031/4.60 ± 0.79	0.214 ± 0.096/2.91 ± 0.68
		w/o Swap	0.096 ± 0.027/3.97 ± 0.67	0.192 ± 0.084/1.73 ± 0.35
		w/o SiLU	0.083 ± 0.023/3.49 ± 0.62	0.187 ± 0.082/1.65 ± 0.28
		DMT-Loc (Ours)	**0.073** ± **0.017**/<u>3.25</u> ± **0.56**	**0.173** ± <u>0.061</u>/**1.00** ± **0.11**

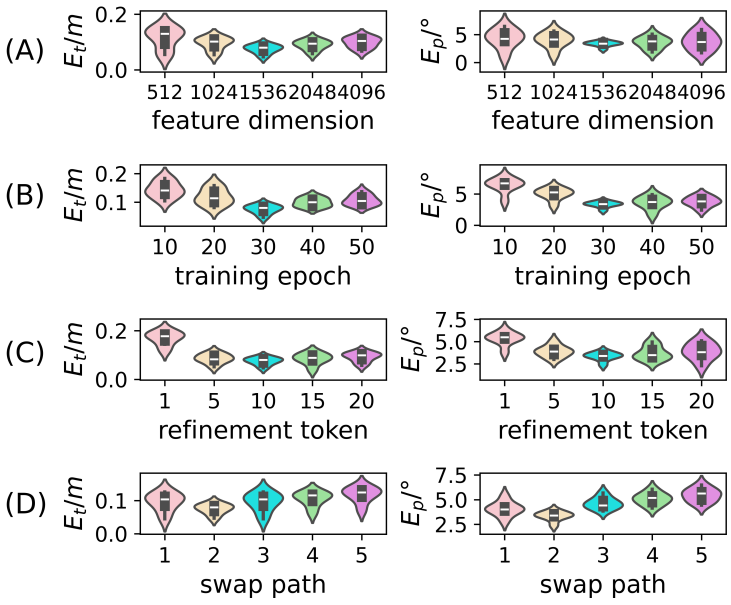


Figure 11. Systematic hyperparameter analysis of DMT-Loc: (**A**) feature dimension; (**B**) training epochs; (**C**) reference tokens; and (**D**) swap paths. The left and right columns reflect translation errors (m) and rotation errors (°), respectively. Optimal settings are highlighted: 1536 dimension, 30 epochs, 10 tokens, and 2 swap paths.

Table 9. Sensitivity analysis of modular hyperparameters. Sensitivity is reported as the maximum relative change in median pose errors (averaged over translation/rotation and both 7Scenes and Cambridge Landmarks datasets), varying each parameter within the tested range while keeping others at default. † Performance plateaus after 6 CMP layers.

Symbol	Description	Equation	Default Value	Tested Range	Sensitivity
γ	Gaussian kernel scale	Equation (2)	1.0	[0.5, 2.0]	<8%
α	Spatial scaling factor 1	Equation (4)	0.5	[0.3, 0.7]	<5%
β	Spatial scaling factor 2	Equation (4)	2.0	[1.5, 2.5]	<5%
ϵ	Stability constant	Equation (7)	1×10^{-4}	$\{10^{-5}, 10^{-3}\}$	<1%
k	Number of HNSW candidates	Equation (8)	10	[5, 20]	<6%
τ	Number of CMP layers †	Equation (9)	6	[2, 10]	<10%
λ	Loss balance coefficient 1	Equation (12)	1	[0.5, 1.5]	<2%
μ	Loss balance coefficient 2	Equation (12)	1	[0.5, 1.5]	<4%

5.4. Latency and Memory

Table 10 outlines the computational overheads of various RPE models processing a query image in 640 × 480 resolution. With superior indoor accuracy, DMT-Loc 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, DMT-Loc reduces pose regression time to just 6.09 ms. In contrast, ReLoc3r incurs a significant memory overhead exceeding 400 GB FLOPs due to implicit stereo computations, whereas DMT-Loc reduces this by 34.61%. These results confirm that DMT-Loc is more suitable for real-time robotic applications where both low latency and high precision are required.

Table 10. Latency and memory for online RPR inference on a single image within 640 × 480 resolution. "↓" indicates that lower values are better. (FLOPs: floating point operations). Best results are highlighted in bold and second best are underlined.

Method	Extraction Latency (ms) ↓	Regression Latency (ms) ↓	Params (MB) ↓	FLOPs (GB) ↓
NN-Net [32]	40.19	8.41	**47.32**	**87.6**
ReLocNet [33]	41.80	9.32	<u>75.30</u>	116.4
Relformer [35]	106.47	18.92	291.45	359.8
Map-free [56]	<u>31.71</u>	11.20	87.26	<u>107.1</u>
RelPoseGNN [36]	72.63	28.28	204.41	284.2
ReLoc3r [37]	83.09	135.36	353.89	419.8
DMT-Loc-pw (Ours)	**10.18**	**6.09**	269.17	274.5
DMT-Loc-mv (Ours)	**10.18**	<u>8.27</u>	269.17	274.5

To assess feasibility on edge deployment, we further estimate theoretical throughput on the NVIDIA Jetson AGX Orin. Based on its 275 TOPS INT8 and 204 GB/s bandwidth, DMT-Loc (274.5 GFLOPs) could achieve ∼15–20 FPS, which is sufficient for real-time robotics. Although SSM kernels are currently less mature than optimized CNNs, recent advances (e.g., official Mamba CUDA implementation) are rapidly bridging this gap.

6. Conclusions

This paper has presented DMT-Loc, a debiased multiplex tokenizer that reframes MFVR as a unified process of invariant feature aggregation and autoregressive anchor interaction. Beyond achieving SoTA accuracy and efficiency, our work establishes that explicit causal modeling of visual–geometric relationships is key to bridging discrete retrieval and continuous pose estimation. The synergistic integration of CMP and SSG-MLP offers a principled and validated MFVR framework.

Limitations: The proposed DMT-Loc highly depends on pre-trained 2D features in challenging conditions, while its causal modeling is applied solely during token matching.

Despite demonstrating cross-scene generalization on held-out datasets, its zero-shot performance on radically heterogeneous domains remains to be validated. Extending causal reasoning throughout the feature encoding pipeline represents a significant avenue for improving robustness and interpretability.

Future work will focus on novel view synthesis for multi-view rendering to mitigate viewpoint redundancy and enhance model generalization. Moreover, it is also critical to validate DMT-Loc on physical edge devices (e.g., NVIDIA Jetson) and advance efficient SSM implementations for resource-constrained platforms, toward deploying robust spatial intelligence in real-world environments.

Author Contributions: Conceptualization, W.W. and H.L.; methodology, W.W. and H.L.; software, W.W. and D.C.; validation, W.W. and P.J.; formal analysis, W.W. and H.L.; investigation, W.W. and P.J.; resources, H.L., S.L. and R.D.; data curation, S.L. and R.D.; writing—original draft preparation, W.W.; writing—review and editing, W.W. and H.L.; visualization, W.W. and D.C.; supervision, H.L.; project administration, H.L. and R.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was jointly supported by grants from the National Key Research and Development Program of China (No. 2024YFB4709802), the National Natural Science Foundation of China (No. 62373009), the 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).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are contained within the article. Derived code supporting the findings is available on: https://github.com/wwsbot/DMT-Loc (accessed on 10 March 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MFVR	Map-Free Visual Relocalization
APE	Absolute Pose Estimation
RPE	Relative Pose Estimation
VPR	Visual Place Recognition
RPR	Relative Pose Regression
MxIT	Multiplex Interacive Tokenization
SKSA	Sparse Kernel Self-Attention
TCDA	Triplet Cross-Domain Aggregation
DAR	Debiased Anchor Registration
HNSW	Hierarchical Navigable Small World
CMP	Causal Mamba Pointer
GIPR	Geometry-Informed Pose Regression
SSG-MLP	Symmetric Swap Gated Multi-Layer Perceptron

Appendix A. Encoder Structure and Versions

During the process of feature map encoding, MambaVision [45] is taken as the visual backbone to yield a preliminary feature map for each scene image. The pre-trained encoder is shown in Figure A1. It is a hybrid backbone that combines convolutional blocks with Mamba blocks effectively to perform local feature extraction and capture long-range spatial dependencies, respectively.

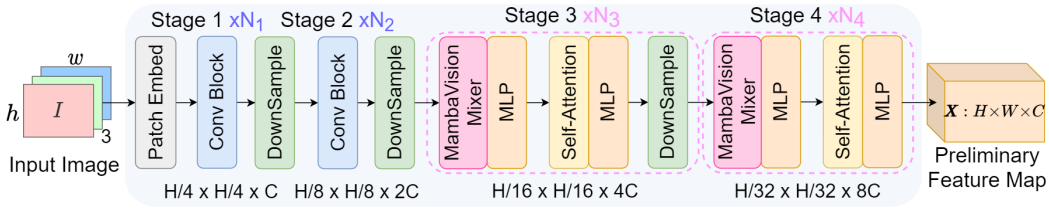


Figure A1. The pipeline of pre-trained visual backbone for DMT-Loc: MambaVision [45]. There are four stages with the first two convolution blocks and the last two Mamba blocks, while the needless average pooling layer and fully connected layer are omitted.

Concretely, MambaVision comes in three versions as summarized in Table A1: tiny (MambaVison-T), base (MambaVison-B), and large (MambaVison-L). The large version achieves the best performance at the cost of greater overhead. To understand the impact of pre-training, MambaVision-L pre-trained on ImageNet-1K is also compared against the default 21K version, demonstrating that larger-scale pre-training unlocks additional potential of the Mamba architecture for visual representation.

Table A1. Model overheads and median errors of different versions for the pretrained MambaVision on 7Scenes and Cambridge Landmarks datasets. "↓" indicates that lower values are better. Best RPR results are highlighted in bold and second best are underlined.

Model	Pretrain	Params (MB) ↓	FLOPs (GB) ↓	7Scenes (cm/°) ↓	Cambridge (cm/°) ↓
MambaVision-T	ImageNet-21K	**31.8**	**4.4**	8.6/4.31	16.9/1.57
MambaVision-B	ImageNet-21K	<u>97.7</u>	<u>15.0</u>	7.9/3.83	16.1/1.39
MambaVision-L	ImageNet-1K	241.5	37.5	<u>7.7</u>/<u>3.58</u>	<u>15.8</u>/<u>1.26</u>
MambaVision-L	ImageNet-21K	241.5	37.5	**7.4**/**3.25**	**15.5**/**1.02**

Appendix B. Token and Error Distributions

To analyze the discriminative power of the learned tokens, we visualize their distributions using t-SNE on the 7Scenes and Cambridge Landmarks datasets. As shown in Figure A2, tokens generated by DMT-Loc form more compact and well-separated clusters than those from NetVLAD, MixVPR, and SelaVPR. This clear structural separation indicates stronger discriminative power and aligns with DMT-Loc's superior generalization performance in scene-agnostic evaluations (Section 5.2.1). The well-formed token distribution captures invariant scene characteristics with minimal inter-scene overlap, providing a robust foundation for both image retrieval and pose regression, and further supports the efficiency and versatility of DMT-Loc as demonstrated in the main experiments.

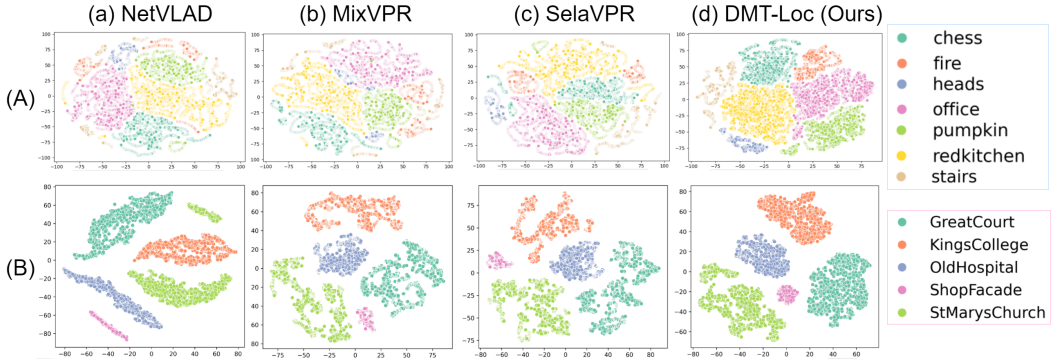


Figure A2. t-SNE visualizations of token distributions on (**A**) 7Scenes and (**B**) Cambridge Landmarks datasets, corresponding to: (**a**) NetVLAD; (**b**) MixVPR; (**c**) SelaVPR; and (**d**) our proposed DMT-Loc.

To further validate the robustness of our debiased design, the full distribution of pose errors is analyzed beyond median statistics. Table A2 reports the 90th and 95th percentile errors of DMT-Loc compared with competitive baselines on the most challenging scenes

(i.e., 7Scenes Stairs and Cambridge Church) as well as the average over all scenes of each dataset. DMT-Loc consistently achieves lower tail errors, e.g., reducing the 95th percentile translation error on the Stairs scene by 46.4% (0.15 m vs. 0.28 m) and rotation error on the Church scene from 1.8° to 0.9°. These results demonstrate that DMT-Loc not only improves average performance but also effectively mitigates large errors in difficult cases, confirming the robustness of its feature-level debiasing across diverse environments.

Table A2. The 90th and 95th percentile pose errors (cm/°) on representative scenes and overall averages of 7Scenes and Cambridge Landmarks datasets. "↓" indicates that lower values are better. Best RPR results are highlighted in bold and second best are underlined.

Method	7Scenes (Stairs) ↓		7Scenes (Average) ↓		Cambridge (Church) ↓		Cambridge (Average) ↓	
	P90	P95	P90	P95	P90	P95	P90	P95
Map-free [56]	31/4.2	35/5.1	26.3/3.2	31.9/4.6	42/3.8	51/4.9	30.8/3.7	35.5/3.6
RelFormer [35]	38/5.6	44/6.8	23.7/3.6	29.6/4.5	31/2.9	39/3.7	27.8/4.3	33.8/5.2
ReLoc3r [37]	<u>24</u>/<u>3.8</u>	<u>28</u>/<u>4.5</u>	<u>14.2</u>/<u>1.9</u>	<u>17.5</u>/<u>2.4</u>	<u>18</u>/<u>1.2</u>	<u>22</u>/<u>1.8</u>	<u>18.8</u>/<u>2.4</u>	<u>21.7</u>/<u>2.9</u>
DMT-Loc (Ours)	**12/2.1**	**15/2.8**	**7.1/1.2**	**9.3/1.5**	**9/0.7**	**12/0.9**	**9.4/1.4**	**13.1/2.0**

References

1. Kim, J.; Kim, Y.M.; Zahreddine, R.; Welge, W.A.; Krishnan, G.; Ma, S.; Wang, J. Privacy-preserving visual localization with event cameras. *IEEE Trans. Image Process.* **2025**, *34*, 6215–6230. [CrossRef] [PubMed]
2. Miao, J.; Jiang, K.; Wen, T.; Wang, Y.; Jia, P.; Wijaya, B.; Zhao, X.; Cheng, Q.; Xiao, Z.; Huang, J.; et al. A survey on monocular re-Localization: From the perspective of scene map representation. *IEEE Trans. Intell. Veh.* **2025**, *10*, 2519–2550. [CrossRef]
3. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *International Conference on Machine Learning*; PMLR; Association for Computing Machinery: Vienna, Austria, 2024; pp. 62429–62442.
4. Cho, J.; Kim, J.; Kim, J.; Kim, M.; Kang, M.; Hong, S.; Oh, T.H.; Yu, Y. DisCoRD: Discrete tokens to continuous motion via rectified flow decoding. In *IEEE International Conference on Computer Vision*; IEEE: Honolulu, HI, USA, 2025; pp. 14602–14612.
5. Sui, Y.; Mao, W.; Wang, S.; Wang, X.; Wu, J.; He, X.; Chua, T.S. Enhancing out-of-distribution generalization on graphs via causal attention learning. *ACM Trans. Knowl. Discov. Data* **2024**, *18*, 1–24. [CrossRef]
6. Pearl, J. Causal Inference: History, perspectives, adventures, and unification (an interview with Judea Pearl). *Obs. Stud.* **2022**, *8*, 23–36. [CrossRef]
7. Zeng, K.; Lin, H.; Yan, Z.; Fang, J.; Lai, T. Non-local self-attention network for image super-resolution. *Appl. Intell.* **2024**, *54*, 5336–5352. [CrossRef]
8. Malkov, Y.A.; Yashunin, D.A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 824–836. [CrossRef] [PubMed]
9. Shotton, J.; Glocker, B.; Zach, C.; Izadi, S.; Criminisi, A.; Fitzgibbon, A. Scene coordinate regression forests for camera relocalization in RGB-D images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Portland, OR, USA, 2013; pp. 2930–2937.
10. Kendall, A.; Grimes, M.; Cipolla, R. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In *IEEE International Conference on Computer Vision*; IEEE: Santiago, Chile, 2015; pp. 2938–2946.
11. Liu, X.; Zhang, Y. Matrices over quaternion algebras. In *Matrix and Operator Equations and Applications*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 139–183. [CrossRef]
12. Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Las Vegas, NV, USA, 2016; pp. 5297–5307.
13. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: San Francisco, CA, USA, 2010; pp. 3304–3311.
14. Yang, M.; He, D.; Fan, M.; Shi, B.; Xue, X.; Li, F.; Ding, E.; Huang, J. DOLG: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *IEEE International Conference on Computer Vision*; IEEE: Montreal, QC, Canada, 2021; pp. 11772–11781.
15. Hausler, S.; Garg, S.; Xu, M.; Milford, M.; Fischer, T. Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2021; pp. 14141–14152.
16. Wang, R.; Shen, Y.; Zuo, W.; Zhou, S.; Zheng, N. TransVPR: Transformer-based place recognition with multi-level attention aggregation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New Orleans, LA, USA, 2022; pp. 13648–13657.

17. Berton, G.; Masone, C.; Caputo, B. Rethinking visual geo-localization for large-scale applications. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New Orleans, LA, USA, 2022; pp. 4878–4888.
18. Ali-Bey, A.; Chaib-Draa, B.; Giguere, P. MixVPR: Feature mixing for visual place recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision*; IEEE: Waikoloa, HI, USA, 2023; pp. 2998–3007.
19. Izquierdo, S.; Civera, J. Optimal transport aggregation for visual place recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Seattle, WA, USA, 2024; pp. 17658–17668.
20. Lu, F.; Zhang, L.; Lan, X.; Dong, S.; Wang, Y.; Yuan, C. Towards seamless adaptation of pre-trained models for visual place recognition. In *International Conference on Learning Representations*; ICLR: Vienna, Austria, 2024; pp. 1–22.
21. Tzachor, I.; Lerner, B.; Levy, M.; Green, M.; Shalev, T.B.; Habib, G.; Samuel, D.; Zailer, N.K.; Shimshi, O.; Darshan, N.; et al. EffoVPR: Effective foundation model utilization for visual place recognition. In *International Conference on Learning Representations*; ICLR: Singapore, 2025; pp. 1–23.
22. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning robust visual features without supervision. *arXiv* **2023**, arXiv:2304.07193. [CrossRef]
23. Sattler, T.; Leibe, B.; Kobbelt, L. Efficient and effective prioritized matching for large-scale image-based localization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1744–1756. [CrossRef] [PubMed]
24. Giang, K.T.; Song, S.; Jo, S. Learning to produce semi-dense correspondences for visual localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Seattle, WA, USA, 2024; pp. 19468–19478.
25. Brachmann, E.; Rother, C. Visual camera re-localization from RGB and RGB-D images using DSAC. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 5847–5865. [CrossRef] [PubMed]
26. Brachmann, E.; Cavallari, T.; Prisacariu, V.A. Accelerated coordinate encoding: Learning to relocalize in minutes using RGB and poses. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Vancouver, BC, Canada, 2023; pp. 5044–5053.
27. Wang, K.; Jiang, Z.; Dai, K.; Xie, T.; Jin, D.; Li, R.; Zhao, L.; Chen, X. SACNet: A scattered attention-based network with feature compensator for visual localization. *IEEE Robot. Autom. Lett.* **2024**, *9*, 3586–3593. [CrossRef]
28. Moreau, A.; Piasco, N.; Tsishkou, D.; Stanciulescu, B.; de La Fortelle, A. Lens: Localization enhanced by nerf synthesis. In *Conference on Robot Learning*; PMLR; Association for Computing Machinery: Auckland, New Zealand, 2022; pp. 1347–1356.
29. Chen, S.; Cavallari, T.; Prisacariu, V.A.; Brachmann, E. Map-relative pose regression for visual re-localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Seattle, WA, USA, 2024; pp. 20665–20674.
30. Sarlin, P.E.; Cadena, C.; Siegwart, R.; Dymczyk, M. From coarse to fine: Robust hierarchical localization at large scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Los Angeles, CA, USA, 2019; pp. 12716–12725.
31. Saha, S.; Varma, G.; Jawahar, C.V. Improved visual relocalization by discovering anchor points. In *British Machine Vision Conference*; BMVA Press: Newcastle, UK, 2018.
32. Laskar, Z.; Melekhov, I.; Kalia, S.; Kannala, J. Camera relocalization by computing pairwise relative poses using convolutional neural network. In *IEEE International Conference on Computer Vision Workshops*; IEEE: Venice, Italy, 2017; pp. 929–938.
33. Balntas, V.; Li, S.; Prisacariu, V. RelocNet: Continuous metric learning relocalisation using neural nets. In *European Conference on Computer Vision*; Springer: Munich, Germany, 2018; pp. 751–767.
34. Zhou, Q.; Sattler, T.; Pollefeys, M.; Leal-Taixe, L. To learn or not to learn: Visual localization from essential matrices. In *IEEE International Conference on Robotics and Automation*; IEEE: Paris, France, 2020; pp. 3319–3326.
35. Idan, O.; Shavit, Y.; Keller, Y. Beyond familiar landscapes: Exploring the limits of relative pose regressors in new environments. *Comput. Vis. Image Underst.* **2024**, *264*, 104629. [CrossRef]
36. Turkoglu, M.O.; Brachmann, E.; Schindler, K.; Brostow, G.J.; Monszpart, A. Visual camera re-localization using graph neural networks and relative pose supervision. In *IEEE International Conference on 3D Vision*; IEEE: London, UK, 2021; pp. 145–155.
37. Dong, S.; Wang, S.; Liu, S.; Cai, L.; Fan, Q.; Kannala, J.; Yang, Y. 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*; IEEE: Nashville, TN, USA, 2025; pp. 16739–16752.
38. Wang, S.; Leroy, V.; Cabon, Y.; Chidlovskii, B.; Revaud, J. Dust3r: Geometric 3D vision made easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Seattle, WA, USA, 2024; pp. 20697–20709.
39. Martins, A.; Astudillo, R. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*; PMLR; Association for Computing Machinery: New York, NY, USA, 2016; pp. 1614–1623.
40. Zhang, X.; Ding, H.; Xie, F.; Pan, L.; Zi, Y.; Wang, K.; Zhang, H. Beyond spatial domain: Cross-domain promoted fourier convolution helps single image dehazing. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 10221–10229. [CrossRef]
41. Taira, H.; Okutomi, M.; Sattler, T.; Cimpoi, M.; Pollefeys, M.; Sivic, J.; Pajdla, T.; Torii, A. InLoc: Indoor visual localization with dense matching and view synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Salt Lake City, UT, USA, 2018; pp. 7199–7209.
42. Toft, C.; Maddern, W.; Torii, A.; Hammarstrand, L.; Stenborg, E.; Safari, D.; Okutomi, M.; Pollefeys, M.; Sivic, J.; Pajdla, T.; et al. Long-term visual localization revisited. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2074–2088. [CrossRef] [PubMed]

43. Warburg, F.; Hauberg, S.; Lopez-Antequera, M.; Gargallo, P.; Kuang, Y.; Civera, J. Mapillary street-level sequences: A dataset for lifelong place recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Seattle, WA, USA, 2020; pp. 2626–2635.
44. Sünderhauf, N.; Neubert, P.; Protzel, P. Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons. In *IEEE International Conference on Robotics and Automation*; IEEE: Karlsruhe, Germany, 2013; p. 2013.
45. Hatamizadeh, A.; Kautz, J. MambaVision: A hybrid mamba-transformer vision backbone. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Nashville, TN, USA, 2025; pp. 25261–25270.
46. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Miami, FL, USA, 2009; pp. 248–255.
47. Amirian, M.; Schwenker, F. Radial basis function networks for convolutional neural networks to learn similarity distance metric and improve interpretability. *IEEE Access* **2020**, *8*, 123087–123097. [CrossRef]
48. Sarlin, P.E.; Unagar, A.; Larsson, M.; Germain, H.; Toft, C.; Larsson, V.; Pollefeys, M.; Lepetit, V.; Hammarstrand, L.; Kahl, F.; et al. Back to the feature: Learning robust camera localization from pixels to pose. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2021; pp. 3247–3257.
49. Polizzi, V.; Cannici, M.; Scaramuzza, D.; Kelly, J. FaVoR: Features via voxel rendering for camera relocalization. In *IEEE/CVF Winter Conference on Applications of Computer Vision*; IEEE: Tucson, AZ, USA, 2025; pp. 44–53.
50. Tang, S.; Tang, S.; Tagliasacchi, A.; Tan, P.; Furukawa, Y. Neumap: Neural coordinate mapping by auto-transdecoder for camera localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Vancouver, BC, Canada, 2023; pp. 929–939.
51. Wang, S.; Laskar, Z.; Melekhov, I.; Li, X.; Zhao, Y.; Tolias, G.; Kannala, J. HSCNet++: Hierarchical scene coordinate classification and regression for visual localization with transformer. *Int. J. Comput. Vis.* **2024**, *132*, 2530–2550. [CrossRef]
52. Bui, B.T.; Bui, H.H.; Tran, D.T.; Lee, J.H. D2S: Representing sparse descriptors and 3D coordinates for camera relocalization. *IEEE Robot. Autom. Lett.* **2024**, *9*, 11449–11456. [CrossRef]
53. Shavit, Y.; Keller, Y. Camera pose auto-encoders for improving pose regression. In *European Conference on Computer Vision*; Springer: Tel Aviv, Israel, 2022; pp. 140–157.
54. Chen, S.; Li, X.; Wang, Z.; Prisacariu, V.A. Dfnet: Enhance absolute pose regression with direct feature matching. In *European Conference on Computer Vision*; Springer: Tel Aviv, Israel, 2022; pp. 1–17.
55. Lin, J.; Gu, J.; Wu, B.; Fan, L.; Chen, R.; Liu, L.; Ye, J. Learning neural volumetric pose features for camera localization. In *European Conference on Computer Vision*; Springer: MiCo, Milano, 2024; pp. 198–214.
56. Arnold, E.; Wynn, J.; Vicente, S.; Garcia-Hernando, G.; Monszpart, A.; Prisacariu, V.; Turmukhambetov, D.; Brachmann, E. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision*; Springer: Tel Aviv, Israel, 2022; pp. 690–708. [CrossRef]
57. Ding, M.; Wang, Z.; Sun, J.; Shi, J.; Luo, P. CamNet: Coarse-to-fine retrieval for camera re-localization. In *IEEE International Conference on Computer Vision*; IEEE: Seoul, South Korea, 2019; pp. 2871–2880.
58. Sattler, T.; Zhou, Q.; Pollefeys, M.; Leal-Taixe, L. Understanding the limitations of CNN-based absolute camera pose regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Long Beach, CA, USA, 2019; pp. 3297–3307.
59. Dusmanu, M.; Rocco, I.; Pajdla, T.; Pollefeys, M.; Sivic, J.; Torii, A.; Sattler, T. D2-net: A trainable CNN for joint description and detection of local features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Long Beach, CA, USA, 2019; pp. 8092–8101.
60. Ram, P.; Sinha, K. Revisiting kd-tree for nearest neighbor search. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: Anchorage, AK, USA, 2019; pp. 1378–1388.
61. Slaney, M.; Casey, M. Locality-sensitive hashing for finding nearest neighbors. *IEEE Signal Process. Mag.* **2008**, *25*, 128–131. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.