MT-PCR: Hybrid Mamba-Transformer Network with Spatial Serialization for Point Cloud Registration

Bingxi Liu[1,3,*], An Liu[2,*], Hao Chen[4], Huaqi Tao[1], Jinqiang Cui[3], Yiqun Wang[2,†], Hong Zhang[1,†]

[1]Southern University of Science and Technology, Shenzhen, China

[2]Chongqing University, Chongqing, China

[3]Pengcheng Laboratory, Shenzhen, China

[4]University of Cambridge, Cambridge, United Kingdom

Abstract

Point cloud registration (PCR) is a fundamental task in 3D computer vision and robotics. Most learning-based PCR methods rely on Transformer architectures, which suffer from quadratic computational complexity. This limitation restricts the resolution of point clouds that can be processed, inevitably leading to information loss. In contrast, Mamba, a recently proposed model based on state-space models, achieves linear computational complexity while maintaining strong long-range contextual modeling capabilities. However, directly applying Mamba to PCR tasks yields suboptimal performance due to the unordered and irregular nature of point cloud data. To address these challenges, we propose MT-PCR, the first point cloud registration framework that integrates Mamba and Transformer modules. Specifically, we serialize point cloud features using Z-order space-filling curves to enforce spatial locality, enabling Mamba to better model the geometric structure of the inputs. Additionally, we remove the order-indicator module commonly used in Mamba-based sequence modeling, leading to improved performance in our setting. The serialized features are then processed by an optimized Mamba encoder, followed by a Transformer-based feature refinement stage. Extensive experiments on multiple benchmarks demonstrate that MT-PCR outperforms Transformer-based and other state-of-the-art methods in both accuracy and efficiency, significantly reducing GPU memory usage and FLOPs.

Keywords: Point Cloud Registration, Attention Model, State-Space Model, Spatial Serialization.

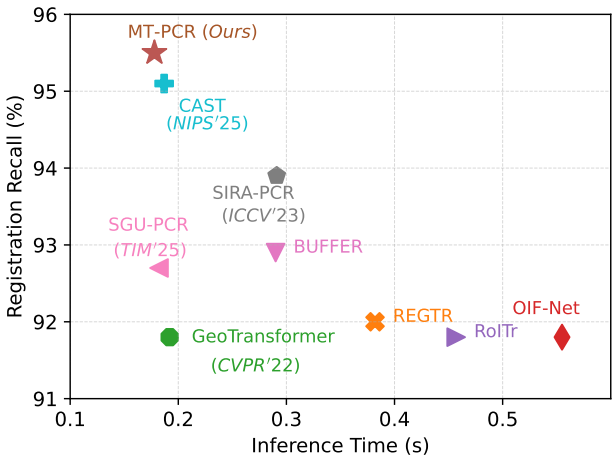

Figure 1. **Registration recall and inference time comparison on 3DMatch.** Our method, MT-PCR, achieves the best registration performance while maintaining competitive inference efficiency, outperforming recent state-of-the-art methods such as CAST (NIPS'25) and SGU-PCR (TIM'25).

1. Introduction

Point cloud registration (PCR) is a foundational task in 3D computer vision [41] and robotics [43], widely applied in domains such as 3D measurement [28], simultaneous localization and mapping (SLAM) [5], augmented reality (AR) [6], and autonomous driving [27]. PCR aims to estimate an optimal rigid transformation that aligns partially overlapping 3D point cloud pairs into a unified coordinate system. Despite significant advances, achieving efficient and accurate PCR remains challenging due to complex spatial structures [40], partial overlaps [16], measurement noise, and large-scale scenes [26].

Recently, Transformer-based methods have demonstrated remarkable performance improvements in PCR tasks by leveraging powerful self-attention and cross-attention mechanisms [33]. Transformers [38] effectively model global spatial relationships and feature correspon-

*Equal contribution.

†Corresponding authors.

dences across point clouds, significantly outperforming hand-crafted or convolution-based approaches. However, Transformers inherently exhibit quadratic computational complexity with respect to sequence length [9], severely limiting their scalability and resolution capacity. Consequently, existing Transformer-based methods typically rely on downsampling point clouds to reduce computational overhead, which inevitably leads to the loss of important geometric information critical for precise alignment.

On the other hand, linear attention models, particularly the recently introduced Mamba model [14], have demonstrated promising results in efficiently capturing the long-range contextual dependencies in sequential modeling tasks. Unlike Transformers, Mamba leverages linear-complexity state-space architectures to approximate global context, thereby significantly enhancing computational efficiency and scalability for long sequences. Nevertheless, directly applying the Mamba model to PCR tasks leads to suboptimal registration accuracy, primarily due to the absence of explicit spatial modeling and inadequate handling of unstructured point cloud data.

Motivated by these observations, we propose MT-PCR, *the first* hierarchical point cloud registration framework that combines the efficient sequence modeling capabilities of Mamba with the bidirectional spatial awareness enabled by cross-attention modules. To effectively adapt Mamba to irregular point cloud data, we introduce a feature serialization strategy based on Z-order space-filling curves, which enforces spatial locality and enhances compatibility between point cloud structures and Mamba's sequence modeling paradigm. Furthermore, we observe that removing the order-indicator tokens, which are typically employed in Mamba for sequential tasks, improves registration performance on 3D data.

Extensive experiments on standard PCR benchmarks, including the widely used 3DMatch [49], 3DLoMatch [19], KITTI [12], and ETH-Challenges [30] datasets, demonstrate that MT-PCR significantly outperforms existing Transformer-based methods and other state-of-the-art (SOTA) approaches, as illustrated in Fig. 1. Moreover, our method achieves these improvements while drastically reducing GPU memory usage and computational overhead in terms of FLOPs, thereby validating the scalability and efficiency of our hybrid architecture, as shown in Fig. 2.

In summary, *our contributions* encompass four key aspects as follows:

- We introduce MT-PCR, the ***first*** hybrid Mamba-Transformer framework for hierarchical point cloud registration, which leverages linear-complexity sequence modeling alongside bidirectional cross-attention mechanisms.
- We propose a spatially aware serialization method based on Z-order space-filling curves, enabling Mamba to effectively process unstructured point cloud data in a sequential format.
- We demonstrate that the order-indicator module, commonly adopted in Mamba-based sequence modeling, is unnecessary in our context, and its removal results in improved performance on PCR tasks.
- We conduct comprehensive experiments across multiple benchmarks, showing that MT-PCR achieves state-of-the-art performance while significantly enhancing computational efficiency in terms of memory usage and FLOPs.

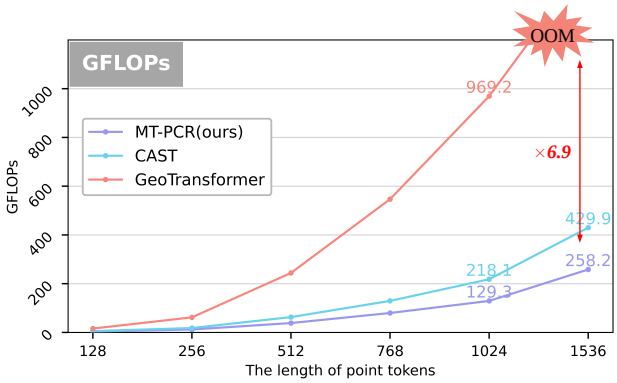

Figure 2. **FLOPs comparison under varying point token lengths.** MT-PCR scales significantly better than GeoTransformer and CAST, maintaining low computational overhead even as the input size increases. Notably, GeoTransformer suffers from *out-of-memory* (OOM) issues at large resolutions, while MT-PCR remains efficient, achieving up to $6.9\times$ lower FLOPs at 1536 tokens.

2. Related Work

2.1. Point Cloud Registration

PCR aims to estimate a rigid transformation between point clouds. Early methods primarily adopted hand-crafted descriptors to represent local features, typically leveraging local geometric attributes [37] or spatial distribution histograms [35, 36]. Nonetheless, the representational capacity of hand-crafted features is inherently limited, often leading to matching failures in noisy or complex scenarios.

Recently, learning-based 3D descriptors have demonstrated significant advantages. PerfectMatch [13] employs a Smoothed Density Value representation to learn discriminative features. PPF [11] extracts globally context-aware, patch-wise features using a PointNet [32]-based architecture. FCGF [10] utilizes a sparse 3D convolutional encoder–decoder network for dense descriptor learning. SpinNet [2] proposes a 3D cylindrical convolutional network with specialized coordinate-system designs to achieve

rotation invariance. Predator [19] integrates graph convolutions with cross-attention mechanisms, explicitly optimizing overlap-region prediction for low-overlap scenarios.

Inspired by Transformer architectures, recent studies have incorporated attention mechanisms into PCR networks to enhance robustness. CoFiNet [46] pioneers the integration of self-attention and cross-attention modules for coarse-level feature matching, combined with optimal transport theory to refine fine-grained correspondences. GeoTransformer [33] introduces geometry-enhanced self-attention and a local-to-global registration framework to ensure pose-estimation consistency. RoITr [47] develops a rotation-invariant Transformer architecture based on point-pair features, strengthening the robustness of coarse-to-fine frameworks. DiffusionPCR [8] incorporates diffusion models to enable iterative refinement of feature matching. CAST [18] integrates consistency-aware mechanisms and point-guided strategies to suppress interference from irrelevant regions while enhancing feature matching capabilities. However, these methods are inherently limited by the quadratic computational complexity $\mathcal{O}(N^2)$ of Transformer architectures, which restricts the resolvable point-cloud size; consequently, downsampling high-density point clouds inevitably discards geometric detail.

2.2. State Space Models and Mamba

State Space Models (SSMs) [20] originate from control theory and have garnered attention as efficient alternatives to Transformers for sequence modeling. Classical SSMs such as S4 [15] demonstrate the ability to model long-range dependencies with linear complexity by leveraging diagonal-plus-low-rank state transitions and HiPPO initialization. These models often suffer from computational inefficiencies or limited scalability.

To address these issues, Mamba [14] introduces a selective SSM mechanism, where input-conditioned parameterization enables selective information flow. By combining input-dependent state updates with hardware-aware parallelization, Mamba achieves strong performance while maintaining linear-time inference. Variants such as Vision Mamba [51] and Vmamba [24] adapt Mamba to image-level tasks through bidirectional and cross-selective scanning strategies, showcasing its potential as a backbone for visual understanding.

Recent works attempt to transfer the benefits of SSMs and Mamba to 3D point cloud data. MetaLA [9] integrates Mamba into a meta-sequential attention framework, emphasizing lightweight modeling for point cloud perception tasks. PointMamba [23, 22] employs an octree-based serialization scheme to impose order on unordered point cloud data, enabling causal dependencies compatible with Mamba's design. Mamba3D [17] extends this line of research by incorporating Mamba into a 3D-aware architecture. It enhances Mamba's spatial awareness through local geometry integration and various pretraining strategies, demonstrating improved scalability and representational capacity on 3D datasets. However, Mamba3D targets object-level recognition and does not explicitly address the unique demands of PCR. In contrast, our work proposes the first Mamba-based architecture specifically tailored for PCR. We also address key challenges in unordered modeling and build a hierarchical framework that combines Mamba's global modeling capabilities with local attention and cross-scale refinement.

3. Problem Definition

Given two partially overlapping 3D point clouds, $X = \{\mathbf{x}_i \in \mathbb{R}^3 \mid i = 1, 2, \ldots, M\}$ and $Y = \{\mathbf{y}_j \in \mathbb{R}^3 \mid j = 1, 2, \ldots, N\}$, which are denoted as the source and target, respectively, the goal of PCR is to estimate an optimal rigid transformation that aligns X with Y. This transformation is characterized by a rotation matrix $\mathbf{R} \in \mathrm{SO}(3)$ and a translation vector $\mathbf{t} \in \mathbb{R}^3$. The solution is typically obtained by minimizing a weighted sum of squared distances between corresponding point pairs in a predicted correspondence set C:

$$\min_{\mathbf{R},\mathbf{t}} \sum_{(\mathbf{x}_k,\mathbf{y}_k)\in C} w_k \left\|\mathbf{R}\mathbf{x}_k + \mathbf{t} - \mathbf{y}_k\right\|_2^2, \tag{1}$$

where w_k denotes the weight associated with each correspondence $(\mathbf{x}_k, \mathbf{y}_k)$.

4. Method

In this section, we first introduce background on Transformers and Mamba in Sec. 4.1. We then provide an overview of our framework, MT-PCR, in Sec. 4.2. The spatial serialization strategy is described in detail in Sec. 4.3, followed by the design of the Mamba encoder in Sec. 4.4. Finally, we introduce the loss functions used for training in Sec. 4.5.

4.1. Preliminaries

Attention Mechanisms and Transformer. Transformers, constructed by stacking self-attention and cross-attention modules, have been widely employed in PCR tasks to effectively model global dependencies and feature correspondences. The self-attention mechanism computes attention weights within the same point set to capture internal feature interactions, while cross-attention identifies correspondences between two distinct point sets. Formally, given queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$, attention is computed as follows:

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{2}$$

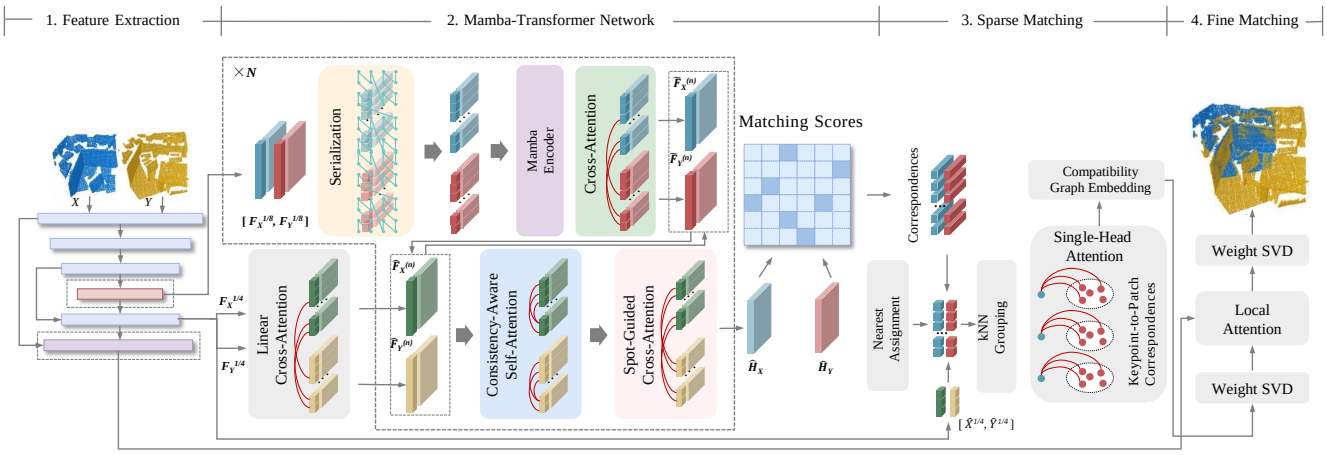

Figure 3. **Overview of the MT-PCR Framework.** The proposed pipeline consists of four stages: multi-scale feature extraction, coarse matching, sparse correspondence refinement, and fine registration. Notably, the coarse matching stage incorporates Mamba encoders with spatial serialization to model global geometric context efficiently.

where d_k denotes the dimensionality of key vectors. Despite their effectiveness, Transformers suffer from quadratic computational complexity with respect to sequence length, limiting their scalability to larger point clouds.

State Space Models and Mamba. SSMs can be viewed as linear time-invariant, multi-input multi-output (MIMO) systems. Mathematically, a continuous-time SSM is described by a set of ordinary differential equations (ODEs):

$$\mathbf{h}'(t) = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t), \tag{3}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{h}(t) + \mathbf{D}\mathbf{x}(t), \tag{4}$$

where $\mathbf{x}(t) \in \mathbb{R}^L$, $\mathbf{h}(t) \in \mathbb{R}^N$, $\mathbf{y}(t) \in \mathbb{R}^L$, and $\mathbf{h}'(t) \in \mathbb{R}^N$ represent the continuous-time input, hidden state, output, and the derivative of the hidden state, respectively. $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the state matrix, $\mathbf{B} \in \mathbb{R}^{N\times L}$ is the input matrix, $\mathbf{C} \in \mathbb{R}^{L\times N}$ is the output matrix, and $\mathbf{D} \in \mathbb{R}^{L\times L}$ is the feed-through matrix.

The continuous-time system can be discretized into a discrete-time SSM via zero-order hold (ZOH) discretization. The parameters $\mathbf{A}$ and $\mathbf{B}$ of the discrete-time SSM can be obtained by introducing the sampling step Δ and applying a Taylor expansion. In this case, the parameters can be approximated as:

$$\overline{\mathbf{A}} = \exp(\Delta\mathbf{A}), \quad \overline{\mathbf{B}} = (\exp(\Delta\mathbf{A}) - \mathbf{I})\,(\Delta\mathbf{A})^{-1} \cdot \Delta\mathbf{B}, \tag{5}$$

which results in the discrete form:

$$\mathbf{h}_k = \overline{\mathbf{A}}\mathbf{h}_{k-1} + \overline{\mathbf{B}}\mathbf{x}_k, \tag{6}$$

$$\mathbf{y}_k = \overline{\mathbf{C}}\mathbf{h}_k + \overline{\mathbf{D}}\mathbf{x}_k, \tag{7}$$

where $\mathbf{x}_k$, $\mathbf{h}_k$, and $\mathbf{y}_k$ represent discrete-time input, state, and output vectors, respectively.

Inspired by the above formulations, Gu and Dao proposed the Mamba model [14], a novel variant of SSMs that introduces input-dependent and time-varying system parameters. Specifically, Δ, $\mathbf{A}$, and $\mathbf{B}$ dynamically adapt according to the input $\mathbf{x}_t$. The Mamba model achieves sequence-modeling performance comparable to Transformers while maintaining linear computational complexity during inference. Although direct parallel computation is challenging, this issue is addressed via a global convolutional expansion:

$$\mathbf{K} = (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, \ldots, \mathbf{C}\overline{\mathbf{A}}^{M-1}\overline{\mathbf{B}}), \quad \mathbf{y} = \mathbf{x} * \mathbf{K}, \tag{8}$$

where M denotes the input sequence length and $\mathbf{K}$ represents the global convolution kernel. This convolutional formulation significantly improves computational efficiency. Further details can be found in the Mamba framework [14].

4.2. MT-PCR Overview

As illustrated in Fig. 3, MT-PCR employs a coarse-to-fine, multi-level feature matching architecture to achieve accurate PCR. It consists of the following key steps:

Multi-scale Feature Extraction: Kernel Point Convolution (KPConv) is utilized to encode input point clouds into multi-scale feature representations spanning from the original dense point cloud (level L_0) to coarser downsampled point clouds (level L_k). The decoder-generated feature maps are denoted as $\mathbf{F}^{1/k} = \{\mathbf{F}_X^{1/k}, \mathbf{F}_Y^{1/k}\}$, corresponding to nodes $X^{1/k}$ and $Y^{1/k}$ downsampled from the original point clouds X and Y, respectively. The key points at the topmost downsampling level are termed superpoints, serving as anchor points for subsequent feature matching.

Mamba-Transformer Coarse Matching: Stacked Mamba feature extraction modules are applied at the superpoint level to construct correspondence sets $\mathcal{C}_s = \{(x_i, y_j)\}$

within overlapping local regions. Based on these correspondences, a differentiable matching matrix $\mathbf{M} \in \mathbb{R}^{N_s \times N_s}$ is constructed via compatibility graph convolutional networks, where each matrix element m_{ij} indicates the matching confidence between superpoint pairs (x_i, y_j).

Sparse Correspondence Refinement: Based on coarse matching results, discriminative keypoints are identified within semi-dense node neighborhoods denoted as $X^{1/4}$ and $Y^{1/4}$. Virtual correspondences for these keypoints are then predicted using a lightweight attention module. To further eliminate spatially inconsistent outliers, a compatibility graph embedding network is employed, yielding a set of high-confidence inliers $\mathcal{I}_{\text{inlier}}$.

Fine Registration. An alignment between the source $\mathbf{X}^{1/2}$ and the target $\mathbf{Y}^{1/2}$ is performed first using the initial transformation $(\hat{\mathbf{R}}_0, \hat{\mathbf{t}}_0)$. Then, a lightweight local attention module is applied to the point sets to estimate refined dense correspondences. These correspondences are then used to compute the final rigid transformation $(\hat{\mathbf{R}}, \hat{\mathbf{t}})$ via weighted SVD.

Challenges and Insights. Replacing the self-attention module with the Mamba module offers a promising pathway toward efficient global modeling due to its linear-time complexity. However, this substitution poses a fundamental challenge: **point clouds are inherently unordered and spatially irregular**, while Mamba is designed for causal, structured sequences as found in natural language processing. Unlike Transformers, which are permutation-invariant and better suited for unordered inputs, Mamba requires well-structured sequential inputs to operate effectively. This discrepancy raises a critical question: *How can we convert 3D point cloud data into meaningful one-dimensional sequences suitable for Mamba while preserving spatial coherence and geometric information?*

To overcome these challenges, we introduce a traversal-based serialization strategy, along with its variants, to generate multiple point-cloud sequences. This approach maximizes the retention of topological associations through multi-path spatial traversal. Additionally, we observe that removing order-indicator tokens, which are typically used in Mamba for sequential tasks, improves registration performance.

4.3. Z-order-based Spatial Serialization

The superpoint features $\mathbf{F}^{1/k}$ obtained from KP-Conv are inherently unordered, which poses challenges for state space models like Mamba that rely on directional, sequential processing. To enable effective sequence modeling, we convert the 3D point cloud into a one-dimensional sequence while preserving spatial locality.

Formally, given a point cloud $\mathcal{X} = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbb{R}^3, i = 1, \ldots, N\}$, we define a bijective serialization function $f : \mathbb{R}^3 \to \mathbb{R}$ that maps each point to a scalar index: $f : \mathbf{x}_i \mapsto s_i, \quad s_i \in \mathbb{R}, \quad \forall \mathbf{x}_i \in \mathcal{X}$.

Algorithm 1 Z-order Encoding for 3D Point Clouds

Require: Grid coordinates $\mathbf{G} = \{(x_i, y_i, z_i)\}_{i=1}^N$, depth d, batch labels $\mathbf{b}$
Ensure: Z-order serialization code $\pi(\mathbf{p}_i)$ for each point

```
for all p_i = (x_i, y_i, z_i) ∈ G do
    Convert x_i, y_i, z_i to d-bit binary integers
    Interleave bits from (x_i, y_i, z_i) to get Morton code m_i
    if batch labels b are provided then
        Concatenate batch ID as prefix:
            m_i ← (batch_i ≪ 3d) | m_i
    end if
end for

return Z-order code list {π(p_i) = m_i}_{i=1}^N
```

Among various strategies, space-filling curves (SFCs) like the Z-order curve (Morton code) are particularly effective, as they preserve spatial locality:

$$\|\mathbf{x}_i - \mathbf{x}_j\|_2 \approx 0 \Rightarrow |s_i - s_j| \approx 0. \tag{9}$$

This ensures that neighboring points in 3D space remain adjacent in the serialized 1D sequence, which is critical for maintaining structural coherence in Mamba-based modeling.

As shown in Alg. 1, we adopt Z-order serialization by first quantizing 3D coordinates to a discrete grid and then interleaving the bits of the d-bit integer representations of (x, y, z) to compute Morton codes. The resulting Z-order indices $\pi(\mathbf{p}_i)$ define a spatially-aware serialization path, enabling Mamba to effectively process the point cloud sequence while preserving geometric proximity.

4.4. Mamba Encoder

After serialization, the feature tokens z are fed into hybrid Mamba–Transformer blocks to extract hierarchical geometric features. Specifically, features from the two inputs are processed sequentially by a Mamba block to capture global context and a Transformer block to refine correspondences. Each Mamba block consists of Layer Normalization (LN), a Selective State Space Model (SelectiveSSM), depth-wise separable convolutions (DW), and residual connections. The architecture is illustrated in Fig. 4, and the forward pass of the l-th block is given by:

$$F'_{l-1} = \text{LN}(F_{l-1}), \tag{10}$$

$$F'_l = \sigma\left(\text{DW}\left(\text{Linear}\left(F'_{l-1}\right)\right)\right), \tag{11}$$

$$F''_l = \sigma\left(\text{Linear}\left(F'_{l-1}\right)\right), \tag{12}$$

$$F_l = \text{Linear}\left(\text{SelectiveSSM}(F'_l) \odot F''_l\right) + F_{l-1}, \tag{13}$$

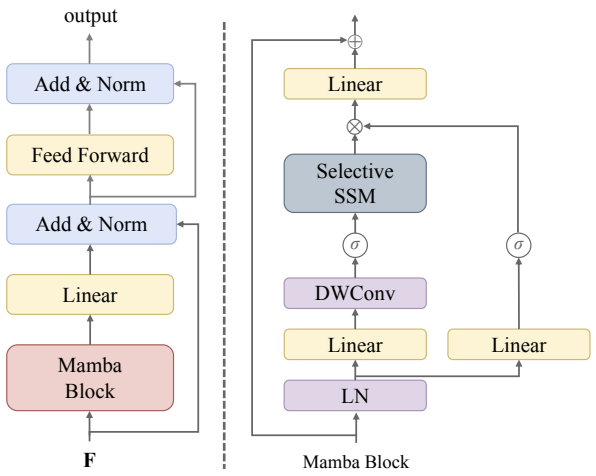

Figure 4. **Architecture of the Mamba Encoder and Block.** The left diagram illustrates the Mamba Encoder with residual connections and feed-forward networks (FNNs). The right diagram shows the internal structure of a Mamba Block, which centers around the SelectiveSSM.

where $F_l \in \mathbb{R}^{2n \times C}$ denotes the output features, and σ represents the SiLU activation function. The SelectiveSSM forms the core of the Mamba block, adaptively modeling contextual dependencies through input-conditioned parameterization. We adopt a minimalist yet effective design [22] that complies with Occam's razor, simplifying the implementation while retaining its expressiveness.

4.5. Loss Functions

Our overall loss function supervises four key modules in the hierarchical registration framework: *keypoint detection*, *coarse matching*, *keypoint matching*, and *dense registration*. Each sub-objective is addressed with a tailored loss component to guide the network effectively at different stages of alignment.

Keypoint Detection. We supervise keypoint detection [21] with $\mathcal{L}_p$. This loss encourages spatial alignment between predicted keypoints from the source and target point clouds while accounting for uncertainty. Specifically, we define:

$$\begin{aligned}\mathcal{L}_p = &\frac{1}{N}\sum_{i=1}^{N}\left(\log\tilde{\sigma}_i + \frac{\|\mathbf{x}_i - \mathbf{y}_{j^*(i)}\|}{\tilde{\sigma}_i}\right) \\ &+ \frac{1}{M}\sum_{j=1}^{M}\left(\log\tilde{\sigma}_j + \frac{\|\mathbf{y}_j - \mathbf{x}_{i^*(j)}\|}{\tilde{\sigma}_j}\right),\end{aligned} \tag{14}$$

where $\mathbf{x}_i \in \mathbb{R}^3$ and $\mathbf{y}_j \in \mathbb{R}^3$ are keypoints from the reference and transformed point clouds, respectively. The indices $j^*(i) = \arg\min_j \|\mathbf{x}_i - \mathbf{y}_j\|$ and $i^*(j) = \arg\min_i \|\mathbf{y}_j - \mathbf{x}_i\|$ denote nearest neighbors. The uncertainty-aware weighting term $\tilde{\sigma}$ is computed as the average predicted variance from both matched keypoints:

$$\tilde{\sigma}_i = \frac{1}{2}\left(\sigma_{x,i} + \sigma_{y,j^*(i)}\right), \quad \tilde{\sigma}_j = \frac{1}{2}\left(\sigma_{y,j} + \sigma_{x,i^*(j)}\right). \tag{15}$$

Coarse Matching. We utilize two weighted cross-entropy losses: the spot-matching loss $\mathcal{L}_s$ for layer-wise coarse scores $\mathbf{P}^{(l)}$ and the coarse-matching loss $\mathcal{L}_c$ for final coarse scores $\mathbf{P}$:

$$\mathcal{L}_s = -\frac{1}{L}\sum_{l=1}^{L}\frac{1}{\sum_{(i,j)\in\mathcal{C}} o_{ij}}\sum_{(i,j)\in\mathcal{C}} o_{ij}\log\mathbf{P}_{ij}^{(l)}, \tag{16}$$

$$\begin{aligned}\mathcal{L}_c = &-\frac{1}{\sum_{(i,j)\in\mathcal{C}} o_{ij}}\sum_{(i,j)\in\mathcal{C}} o_{ij}\log\mathbf{P}_{ij} \\ &-\frac{1}{|\mathcal{N}_X|}\sum_{k\in\mathcal{N}_X}\log(1-\hat{o}_k^X) \\ &-\frac{1}{|\mathcal{N}_Y|}\sum_{k\in\mathcal{N}_Y}\log(1-\hat{o}_k^Y),\end{aligned} \tag{17}$$

where $\mathcal{N}_X$ and $\mathcal{N}_Y$ denote sets of semi-dense nodes in the source and target point clouds without correspondences. The overlap ratio o_{ij} measures how much two local patches overlap. Further details can be found in CAST [18].

Keypoint Matching. We employ three losses to supervise similarity estimation, correspondence prediction, and consistency filtering. First, we use the InfoNCE loss [29], $\mathcal{L}_f$, to maximize the similarity between descriptors d_x and d_{p_x} of true correspondences (x, p_x) and to minimize the similarity between descriptors d_x and d_{n_x} of false correspondences (x, n_x):

$$\mathcal{L}_f = -\mathbb{E}\left[\log\frac{e(d_x^{\mathsf{T}}Wd_{p_x})}{e^{(d_x^{\mathsf{T}}Wd_{p_x})} + \sum_{n_x\in N_x} e(d_x^{\mathsf{T}}Wd_{n_x})}\right], \tag{18}$$

where W is a symmetric learnable weight matrix. Next, we use an ℓ_2 loss, $\mathcal{L}_k$, to supervise predicted correspondences $\hat{\mathbf{y}}$ by minimizing:

$$\mathcal{L}_k = \mathbb{E}_{(x,\hat{y})}\left\|\mathbf{R}x + \mathbf{t} - \hat{y}\right\|_2. \tag{19}$$

Finally, for consistency filtering, we define binary ground-truth labels based on whether a correspondence is an inlier (i.e., its distance is below the threshold R_f) and supervise inlier confidence using binary cross-entropy:

$$\mathcal{L}_i = \text{BCE}(\text{score}, \text{inlier label}). \tag{20}$$

Dense Registration. The dense registration module is supervised using translation and rotation losses:

$$\mathcal{L}_t = \left\|\hat{\mathbf{t}} - \mathbf{t}\right\|_2, \quad \mathcal{L}_R = \left\|\hat{\mathbf{R}}^{\mathsf{T}}\mathbf{R} - \mathbf{I}\right\|_F. \tag{21}$$

The final training loss is formulated as:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_p + \lambda_s \mathcal{L}_s + \lambda_c \mathcal{L}_c + \lambda_f \mathcal{L}_f + \lambda_k \mathcal{L}_k + \lambda_i \mathcal{L}_i \\ &+ \lambda_t \mathcal{L}_t + \lambda_R \mathcal{L}_R,\end{aligned} \tag{22}$$

where $\lambda_s, \lambda_c, \lambda_f, \lambda_k, \lambda_i, \lambda_t, \lambda_R$ are hyperparameters used to balance the contribution of each term by aligning their numerical scales.

5. Evaluation

In this section, we conduct extensive experiments to evaluate the performance of our proposed MT-PCR on both indoor RGB-D datasets (3DMatch [49] and 3DLoMatch [19]) and the outdoor 3D LiDAR dataset (KITTI [12]). To further validate the generalization of the proposed method, we also test it on a 2D LiDAR dataset (ETH-Challenging [31]).

5.1. Implementation Details

We employ the AdamW [25] optimizer, an initial learning rate of 1×10^{-4}, and a weight decay of 1×10^{-4}. A stepwise learning rate scheduler is adopted, decaying the learning rate by 10% every 5 training steps. Gradient clipping with a threshold of 0.5 is used to stabilize the training process. Except for the experiments in Tab. 3, which are performed on an NVIDIA A800 GPU, the rest of the experiments are conducted on a single RTX 3090 GPU. The number of training epochs is set to 5 for 3DMatch and 40 for KITTI. For the KITTI and nuScenes datasets, we set: $\lambda_f = \lambda_i = 1, \lambda_r = 20, \lambda_t = 5, \lambda_s = 0.1, \lambda_c = 0.2, \lambda_k = 1$. For the 3DMatch dataset, we use: $\lambda_c = 1, \lambda_k = 10$.

5.2. Indoor Scenes: 3DMatch & 3DLoMatch

As shown in Tab. 1, we evaluate MT-PCR alongside several SoTA methods on the widely used indoor PCR benchmarks 3DMatch [49] and 3DLoMatch [19], which respectively represent high-overlap (>30%) and low-overlap (10%–30%) scenarios. To ensure fair and consistent comparison of runtime and resource consumption, all methods are re-implemented within a unified PyTorch framework and tested under identical hardware conditions. Our benchmarks include descriptor-based feature matching approaches as well as non-iterative correspondence-based registration methods.

Since MT-PCR relies on sparse keypoint matching, it typically detects around 1000 keypoints. Therefore, we report the Registration Recall (RR) based on 1000 correspondences, aligning with prior work [18]. Notably, despite using only 1000 correspondences, our method outperforms competing dense matching approaches that rely on 1000 or more sampled points. This demonstrates the superior effectiveness of our sparse matching strategy.

On the 3DMatch dataset, our method achieves a new state-of-the-art RR of 95.5%. On the more challenging 3DLoMatch dataset, MT-PCR attains 75.4% RR, outperforming all descriptor-based baselines and non-iterative matching methods. Thanks to the linear attention mechanism enabled by Mamba, our approach also demonstrates superior efficiency, achieving the fastest runtime among all compared methods.

While our RR on 3DLoMatch is slightly lower than that of PEAL [48] and DiffusionPCR [8], these methods incur over *10*× the runtime overhead compared to ours. This favorable efficiency–accuracy trade-off is particularly important for real-world applications with limited computational resources.

5.3. Outdoor Scenes: KITTI Odometry

Tab. 2 summarizes the evaluation of MT-PCR on the KITTI Odometry dataset [12], a standard benchmark for autonomous driving scenarios. We follow the conventional split: sequences 00–05 for training, 06–07 for validation, and 08–10 for testing. To ensure high-quality training samples, point cloud pairs with at least 10 meters spatial separation are selected. Ground-truth transformations are obtained by aligning GPS/IMU trajectories refined via ICP.

We evaluate MT-PCR using Relative Rotation Error (RRE), Relative Translation Error (RTE), and Registration Recall (RR). Our comparisons include SoTA descriptor-based methods such as FCGF [10], D3Feat [4], SpinNet [2], Predator [19], and correspondence-based approaches including CoFiNet [46], GeoTransformer [33], OIF-Net [42], PEAL [48], DiffusionPCR [8], MAC [50], and CAST [18].

MT-PCR achieves SoTA results in both RR and RRE, outperforming the previous best model (DiffusionPCR). Notably, this is ***the first time*** RRE has dropped below 0.2 degrees. This highlights MT-PCR's strong ability to predict accurate rotations. Although its RTE shows slight differences compared to other leading methods, it remains highly competitive and consistent, underscoring robustness in outdoor LiDAR registration.

5.4. Efficiency Study

A core motivation of our work is to address the quadratic time complexity $\mathcal{O}(N^2)$ and high resource consumption inherent in Transformer-based PCRs. To evaluate the efficiency of our design, we conduct a scalability study by varying the serialized point cloud sequence length and measuring performance in terms of FLOPs and memory usage.

We adopt a progressive stress-testing strategy on a 24GB GPU, gradually increasing input sequence length until memory constraints are reached. As shown in Tab. 3, our method achieves superior inference speed and lower GFLOPs across all lengths. Specifically, at a sequence length of 1536, our method reduces computation cost by

Table 1. Evaluation results on indoor RGBD point cloud datasets. **First**, **second**, and **third** best results are color-highlighted.

	Dataset	Source	3DMatch					3DLoMatch				
			Registration Recall (%)					Registration Recall (%)				
	Samples		5000	2500	1000	500	250	5000	2500	1000	500	250
Descriptor-based	PerfectMatch [13]	*CVPR'19*	78.4	76.2	71.4	67.6	50.8	33.0	29.0	23.3	17.0	11.0
	FCGF [10]	*ICCV'19*	85.1	84.7	83.3	81.6	71.4	40.1	41.7	38.2	35.4	26.8
	D3Feat [4]	*CVPR'20*	81.6	84.5	83.4	82.4	77.9	37.2	42.7	46.9	43.8	39.1
	SpinNet [2]	*CVPR'21*	88.6	86.6	85.5	83.5	70.2	59.8	54.9	48.3	39.8	26.8
	YOHO [39]	*MM'22*	90.8	90.3	89.1	88.6	84.5	65.2	65.5	63.2	56.5	48.0
	Predator [19]	*CVPR'21*	89.0	89.9	90.6	88.5	86.6	59.8	61.2	62.4	60.8	58.1
Correspondence-based	REGTR [45]	*CVPR'22*			92.0					64.8		
	GeoTransformer [33]	*TPAMI'23*	92.0	91.8	91.8	91.4	91.2	75.0	74.8	74.2	74.1	73.5
	OIF-Net [42]	*NIPS'22*	92.4	91.9	91.8	92.1	91.2	76.1	75.4	75.1	74.4	73.6
	RoITr [47]	*CVPR'23*	91.9	91.7	91.8	91.4	91.0	74.7	74.8	74.8	74.2	73.6
	PEAL [48]	*CVPR'23*	94.4	94.1	94.1	93.9	93.4	**79.2**	79.0	78.8	78.5	77.9
	BUFFER [1]	*CVPR'23*			92.9					71.8		
	SIRA-PCR [7]	*ICCV'23*	93.6	93.9	93.9	92.7	92.4	73.5	73.9	73.0	73.4	71.1
	DiffusionPCR [8]	*arXiv'24*	94.4	94.3	**94.5**	94.0	93.9	80.0	**80.4**	79.2	78.8	78.8
	CAST [18]	*NIPS'25*			**95.2**					75.1		
	SGU-PCR [18]	*TIM'25*	93.4	92.8	92.7	92.4	91.5	75.4	74.9	74.3	74.1	74.0
	MT-PCR	*Ours*			**95.5**					**75.4**		

Table 2. Registration performance on KITTI dataset. **First**, **second**, and **third** best results are color-highlighted.

Model	RTE (cm)	RRE (°)	RR (%)
3DFeat-Net [44]	25.9	0.25	96.0
FCGF [10]	9.5	0.30	96.6
D3Feat [4]	7.2	0.30	99.8
SpinNet [2]	9.9	0.47	99.1
Predator [19]	6.8	0.27	99.8
CoFiNet [46]	8.2	0.41	99.8
GeoTrans [33]	6.8	0.24	99.8
OIF-Net [42]	6.5	**0.23**	99.8
PEAL [48]	6.8	**0.23**	99.8
DiffusionPCR [8]	**6.3**	**0.23**	99.8
MAC [50]	8.5	0.40	99.5
CAST [18]	**2.5**	**0.27**	**100.0**
MT-PCR	**2.6**	**0.16**	**100.0**

85.4% (to 1/6.9 GFLOPs) and memory usage by 58.3% (to 1/2.4) compared to a Transformer baseline—while maintaining comparable or better performance. This highlights the significant computational advantage of our model for long sequence modeling.

5.5. Visualization

Fig. 5 provides qualitative results on registration performance on 3DMatch. Compared to CAST, MT-PCR produces more accurate alignment results, especially in challenging regions highlighted by red boxes. The scenes registered by MT-PCR are visibly closer to the ground truth, demonstrating superior robustness and geometric consistency.

Table 3. **Efficiency Comparison at Varying Token Lengths.** Computational efficiency for different methods across increasing token lengths. * On an 80GB GPU, GeoTransformer measured a GFLOPs of 1803 and memory usage of 24,614 MB at token length 1536.

Length of Token	128	256	512	768	1024	1536
	GFLOPs ↓					
GeoTrans [33]	16	62	244	547	969	out of memory*
CoFiNet [46]	8	18	58	147	479	out of memory*
CAST [18]	6	18	63	129	218	430
MT-PCR (*ours*)	**4**	**12**	**39**	**80**	**129**	**258**
	GPU Memory (MB) ↓					
GeoTrans [33]	176	634	2463	5510	12335	out of memory*
CoFiNet [46]	147	573	1947	3709	8732	out of memory*
CAST [18]	120	**334**	1185	2491	4189	11028
MT-PCR (*ours*)	**119**	335	**1162**	**2428**	**4091**	**10591**
	FPS ↑					
GeoTrans [33]	5.97	5.43	4.79	3.27	2.83	out of memory*
CoFiNet [46]	5.44	4.68	3.35	2.75	1.68	out of memory*
CAST [18]	**6.52**	**6.19**	**5.27**	4.50	3.69	2.63
MT-PCR (*ours*)	6.17	5.92	5.22	4.59	**4.02**	**3.00**

5.6. Ablation Study

Effectiveness of Architecture. The results in Tab. 4 demonstrate the advantage of the hybrid Mamba-Transformer (HMT) architecture over single backbone models. HMT consistently outperforms both Mamba and Transformer alone in terms of RR, PMR, and PIR, confirming that the combination effectively leverages complementary strengths.

Effectiveness of Mamba. In addition to Mamba, several newly proposed models with linear computational complexity have recently emerged, as shown in Tab. 5. To evalu-

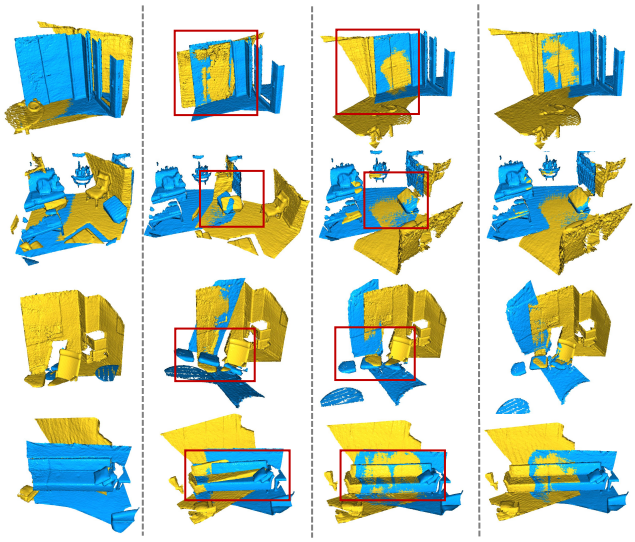

Figure 5. **Qualitative registration results** of CAST and MT-PCR compared with the ground truth alignment on 3DMatch dataset. We present four examples in four rows, which demonstrate the robustness and accuracy of our method.

Table 4. **Comparison of different architectures.**

Method	RR (%)	PMR (%)	PIR (%)
Mamba	92.34	94.17	71.68
Transformer	94.96	95.88	75.54
HMT	**95.54**	**96.87**	**79.65**

Table 5. **Ablation studies on linear attention methods.** The tested linear attention methods are incorporated within our proposed hybrid architecture and improvements. The method names are used here solely for ease of reference.

Model	RR (%)	PMR (%)	PIR (%)
Based [3]	94.75	96.33	78.24
Samba [34]	94.44	96.06	76.46
Metala [9]	94.53	95.36	73.17
Mamba	**95.54**	**96.87**	**79.65**

ate their effectiveness in the context of PCR, we conducted ablation studies comparing Mamba with representative linear attention models. The results suggest that the improved Mamba variant achieves the best performance among these linear models in registration accuracy.

Effectiveness of Serialization. We next examine the impact of different serialization strategies using spatial space-filling curves. In particular, we compare the performance of two representative curves: Hilbert and Z-order, along with their axis-permuted variants: Trans-Hilbert and Trans-Z-order. We also include the commonly used XYZ-order and its variant.

As shown in Tab. 6, methods using space-filling curve

Table 6. **Ablation studies on serialization strategies.**

Strategy	RR (%)	PMR (%)	PIR (%)
baseline	93.47	96.24	76.84
hilbert $\times$ 3	94.36	96.42	77.09
hilbert$^{\mathsf{T}}$ $\times$ 3	94.08	95.97	75.10
hilbert, hilbert$^{\mathsf{T}}$, z	95.07	96.84	78.83
xyz, zxy, yxz	94.00	96.33	76.57
z$^{\mathsf{T}}$ $\times$ 3	94.37	96.51	78.16
z $\times$ 3	**95.54**	**96.87**	**79.65**

Table 7. **Ablation studies on order strategies.**

Setting	RR (%)	PMR (%)	PIR (%)
None	**95.54**	**96.87**	**79.65**
Order indicator	94.74	96.42	77.75

serialization outperform traditional baselines, with the Z-order strategy achieving the best overall performance. We attribute this to the locality-preserving nature of space-filling curves, which provide more logically ordered sequences for state space models. This continuous spatial scanning offers a natural inductive bias for Mamba to model geometry-aware sequences.

Effectiveness of the Order Indicator. In prior work such as PointMamba [22], the Order-Indicator module was necessary because the model used two spatial curves (e.g., Hilbert and T-Hilbert) within the same Mamba encoder to perform bidirectional modeling. Without explicit order distinction, the mixed sequences would interfere with each other in the latent space, degrading performance. In contrast, our method performs all bidirectional information exchange in a subsequent cross-attention module, while the Mamba encoder operates only on a unidirectional sequence generated through dynamic serialization. This design preserves the geometric continuity of the point cloud from the beginning. Introducing an Order-Indicator in our case would artificially break spatial adjacency, weakening Mamba's ability to capture local topology. As shown in Tab. 7, adding the Order-Indicator leads to a 0.8% drop in RR, confirming its negative impact. Therefore, we choose to remove this module in order to maximize the preservation of geometric structure during Mamba processing.

Effectiveness of Mamba and Z-order on Rotation Estimation. As summarized in Tab.8, removing the Mamba encoder significantly degrades RRE from 0.161° to 0.273°, confirming its critical role in capturing global geometric context beyond the backbone. Furthermore, omitting Z-order serialization increases RRE to 0.174°, demonstrating that the spatial locality preserved by Z-order provides a superior inductive bias for high-precision rotation estimation.

Sensitivity of Z-order Depth. We conducted a sensitivity analysis on the depth parameter (d) of the Z-order space-

Table 8. **Ablation Study on Rotation Estimation Components.**

Setting	RR (%)	RRE (°)	PTE (cm)
w/o zorder	**100.0**	0.174	2.64
w/o mamba	**100.0**	0.273	2.56
MT-PCR	**100.0**	**0.161**	**2.55**

filling curve. As shown in Tab. 9, increasing d from 8 to 16 yields only a minor improvement in RR, indicating that the model's performance is not highly sensitive to this parameter.

Table 9. **Sensitivity of Z-order depth parameter d using RR (%).**

Z-order Depth (d)	8	12	16
RR (%)	94.90	95.11	95.54

5.7. Generalization Study

We conduct a generalization experiment by transferring from the outdoor KITTI dataset to another outdoor dataset, ETH-Challenging [31]. Note that the KITTI and ETH-Challenging datasets use Velodyne-64 3D LiDAR and Hokuyo 2D LiDAR sensors, respectively, resulting in significantly different appearances and distributions of point clouds. Hence, this generalization study is both practical for real-world applications and effective in demonstrating the robustness of various methods.

Tab. 10 also presents RRE, RTE, and RR. Our method achieves satisfactory accuracy and robustness, exhibiting superior generalization performance compared to the coarse-to-fine baseline GeoTransformer and other point-wise descriptors. Notably, all point-wise methods show lower registration recall in the generalization setting compared to the patch-wise local descriptor SpinNet [2]. This difference arises primarily because SpinNet employs a feature pyramid network architecture to learn features with abundant global context, which benefits generalization.

Furthermore, we perform an unsupervised domain adaptation (UDA) experiment on MT-PCR, tuning the network to align a point cloud to itself after random rotation and cropping. The results indicate that our model can easily adapt to an unseen domain and achieve robust and accurate performance after only one epoch of unsupervised tuning (approximately 20 minutes on an NVIDIA RTX 3090 GPU).

6. Conclusion

In this paper, we introduced MT-PCR, the first hierarchical point cloud registration framework that unifies the strengths of Mamba and Transformer architectures. By leveraging the linear complexity and long-range modeling capabilities of Mamba, along with the spatially aware refinement of Transformers, our framework addresses the scalability limitations of existing Transformer-based PCR methods. To bridge the gap between sequence-based models and unordered 3D point clouds, we proposed a feature-serialization approach based on Z-order space-filling curves, which preserves spatial locality and improves the compatibility of point-cloud features with Mamba. Additionally, we observed that removing conventional order-indicator tokens, commonly used in sequential modeling, leads to superior registration performance, highlighting a useful design insight for applying state space models to geometric data. Comprehensive experiments on challenging benchmarks such as 3DMatch, 3DLoMatch, KITTI, and the ETH benchmarks demonstrate that MT-PCR not only achieves SoTA accuracy but also significantly improves computational efficiency by reducing both GPU memory consumption and FLOPs. We believe that the hybrid design principles of MT-PCR open up new opportunities for efficient and scalable 3D perception and provide a promising direction for future research in point cloud representation learning and registration.

Table 10. Generalization results on KITTI → ETH datasets.

Method	RTE (cm)	RRE (°)	RR (%)
FCGF	9.08	0.94	45.86
Predator	11.72	1.38	65.64
SpinNet	6.05	0.98	99.44
GeoTransformer	5.97	0.73	91.87
CAST	6.86	0.66	97.05
MT-PCR	6.67	0.67	98.88
MT-PCR + UDA	**4.27**	**0.53**	**99.86**

References

[1] S. Ao, Q. Hu, H. Wang, K. Xu, and Y. Guo. Buffer: Balancing accuracy, efficiency, and generalizability in point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1255–1264, 2023. 8

[2] S. Ao, Q. Hu, B. Yang, A. Markham, and Y. Guo. Spinnet: Learning a general surface descriptor for 3d point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11753–11762, 2021. 2, 7, 8, 10

[3] S. Arora, S. Eyuboglu, M. Zhang, A. Timalsina, S. Alberti, D. Zinsley, J. Zou, A. Rudra, and C. Ré. Simple linear attention language models balance the recall-throughput tradeoff. *ArXiv preprint*, abs/2402.18668, 2024. 9

[4] X. Bai, Z. Luo, L. Zhou, H. Fu, L. Quan, and C.-L. Tai. D3feat: Joint learning of dense detection and description of 3d local features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6359–6367, 2020. 7, 8

[5] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 32(6):1309–1332, 2016. 1

[6] J. Carmigniani, B. Furht, M. Anisetti, P. Ceravolo, E. Damiani, and M. Ivkovic. Augmented reality technologies, systems and applications. *Multimedia tools and applications*, 51:341–377, 2011. 1

[7] S. Chen, H. Xu, R. Li, G. Liu, C.-W. Fu, and S. Liu. Sirapcr: Sim-to-real adaptation for 3d point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14394–14405, 2023. 8

[8] Z. Chen, Y. Ren, T. Zhang, Z. Dang, W. Tao, S. Süsstrunk, and M. Salzmann. Diffusionpcr: Diffusion models for robust multi-step point cloud registration. *arXiv preprint arXiv:2312.03053*, 2023. 3, 7, 8

[9] Y. Chou, M. Yao, K. Wang, Y. Pan, R.-J. Zhu, J. Wu, Y. Zhong, Y. Qiao, B. Xu, and G. Li. Metala: Unified optimal linear approximation to softmax attention map. *Advances in Neural Information Processing Systems*, 37:71034–71067, 2024. 2, 3, 9

[10] C. Choy, J. Park, and V. Koltun. Fully convolutional geometric features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8958–8966, 2019. 2, 7, 8

[11] H. Deng, T. Birdal, and S. Ilic. Ppfnet: Global context aware local features for robust 3d point matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 195–205, 2018. 2

[12] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 2, 7

[13] Z. Gojcic, C. Zhou, J. D. Wegner, and A. Wieser. The perfect match: 3d point cloud matching with smoothed densities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5545–5554, 2019. 2, 8

[14] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2, 3, 4

[15] A. Gu, K. Goel, A. Gupta, and C. Ré. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35:35971–35983, 2022. 3

[16] S. Guo, Y. Wu, B. Xie, B. Liu, and T. Jia. Low-overlap point cloud registration by semiglobal block matching. *IEEE Transactions on Industrial Informatics*, 2024. 1

[17] X. Han, Y. Tang, Z. Wang, and X. Li. Mamba3d: Enhancing local features for 3d point cloud analysis via state space model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4995–5004, 2024. 3

[18] R. Huang, Y. Tang, J. Chen, and L. Li. A consistency-aware spot-guided transformer for versatile and hierarchical point cloud registration. *Proc. Conf. Neural Inf. Process. Syst.*, 2024. 3, 6, 7, 8

[19] S. Huang, Z. Gojcic, M. Usvyatsov, A. Wieser, and K. Schindler. Predator: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4267–4276, 2021. 2, 3, 7, 8

[20] R. E. Kalman. A new approach to linear filtering and prediction problems. 1960. 3

[21] J. Li and G. H. Lee. Usip: Unsupervised stable interest point detection from 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 361–370, 2019. 6

[22] D. Liang, X. Zhou, W. Xu, X. Zhu, Z. Zou, X. Ye, X. Tan, and X. Bai. Pointmamba: A simple state space model for point cloud analysis. In *Advances in Neural Information Processing Systems*, 2024. 3, 6, 9

[23] J. Liu, R. Yu, Y. Wang, Y. Zheng, T. Deng, W. Ye, and H. Wang. Point mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy. *arXiv preprint arXiv:2403.06467*, 2024. 3

[24] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024. 3

[25] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 7

[26] F. Lu, G. Chen, Y. Liu, L. Zhang, S. Qu, S. Liu, R. Gu, and C. Jiang. Hregnet: A hierarchical network for efficient and accurate outdoor lidar point cloud registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[27] W. Lu, Y. Zhou, G. Wan, S. Hou, and S. Song. L3-net: Towards learning based lidar localization for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6389–6398, 2019. 1

[28] J. Luo, D. Dong, and H. Liu. Swift and accurate point cloud registration using sguformer. *IEEE Transactions on Instrumentation and Measurement*, 74:1–12, 2025. 1

[29] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6

[30] F. Pomerleau, M. Liu, F. Colas, and R. Siegwart. Challenging data sets for point cloud registration algorithms. *The International Journal of Robotics Research*, 31(14):1705–1711, Dec. 2012. 2

[31] F. Pomerleau, M. Liu, F. Colas, and R. Siegwart. Challenging data sets for point cloud registration algorithms. *Int. J. Rob. Res.*, 31(14):1705–1711, Dec. 2012. 7, 10

[32] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 2

[33] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, S. Ilic, D. Hu, and K. Xu. Geotransformer: Fast and robust point cloud registration with geometric transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 3, 7, 8

[34] L. Ren, Y. Liu, Y. Lu, Y. Shen, C. Liang, and W. Chen. Samba: Simple hybrid state space models for efficient unlimited context language modeling. *ArXiv preprint*, abs/2406.07522, 2024. 9

[35] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE International Conference on Robotics and Automation*, pages 3212–3217. IEEE, 2009. 2

[36] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz. Aligning point cloud views using persistent feature histograms. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3384–3391. IEEE, 2008. 2

[37] S. Salti, F. Tombari, and L. Di Stefano. Shot: Unique signatures of histograms for surface and texture description. *Computer Vision and Image Understanding*, 125:251–264, 2014. 2

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[39] H. Wang, Y. Liu, Z. Dong, and W. Wang. You only hypothesize once: Point cloud registration with rotation-equivariant descriptors. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1630–1641, 2022. 8

[40] J. Wang and Z. Li. 3dpcp-net: A lightweight progressive 3d correspondence pruning network for accurate and efficient point cloud registration. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1885–1894, 2024. 1

[41] Y. Wang and J. M. Solomon. Deep closest point: Learning representations for point cloud registration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3523–3532, 2019. 1

[42] F. Yang, L. Guo, Z. Chen, and W. Tao. One-inlier is first: Towards efficient position encoding for point cloud registration. *Advances in Neural Information Processing Systems*, 35:6982–6995, 2022. 7, 8

[43] H. Yang, J. Shi, and L. Carlone. Teaser: Fast and certifiable point cloud registration. *IEEE Transactions on Robotics*, 37(2):314–333, 2020. 1

[44] Z. J. Yew and G. H. Lee. 3dfeat-net: Weakly supervised local 3d features for point cloud registration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 607–623, 2018. 8

[45] Z. J. Yew and G. H. Lee. Regtr: End-to-end point cloud correspondences with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6677–6686, 2022. 8

[46] H. Yu, F. Li, M. Saleh, B. Busam, and S. Ilic. Cofinet: Reliable coarse-to-fine correspondences for robust point cloud registration. *Advances in Neural Information Processing Systems*, 34:23872–23884, 2021. 3, 7, 8

[47] H. Yu, Z. Qin, J. Hou, M. Saleh, D. Li, B. Busam, and S. Ilic. Rotation-invariant transformer for point cloud matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5384–5393, 2023. 3, 8

[48] J. Yu, L. Ren, Y. Zhang, W. Zhou, L. Lin, and G. Dai. Peal: Prior-embedded explicit attention learning for low-overlap point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17702–17711, 2023. 7, 8

[49] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1802–1811, 2017. 2, 7

[50] X. Zhang, J. Yang, S. Zhang, and Y. Zhang. 3d registration with maximal cliques. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17745–17754, 2023. 7, 8

[51] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Forty-first International Conference on Machine Learning*, 2024. 3