Dynamic Cues-Assisted Transformer for Robust Point Cloud Registration

Hong Chen, Pei Yan, Sihe Xiang and Yihua Tan*

Hubei Engineering Research Center of Machine Vision and Intelligent Systems,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China
{hongc, yanpei, m202273201, yhtan}@hust.edu.cn

Abstract

Point Cloud Registration is a critical and challenging task in computer vision. Recent advancements have predominantly embraced a coarse-to-fine matching mechanism, with the key to matching the superpoints located in patches with inter-frame consistent structures. However, previous methods still face challenges with ambiguous matching, because the interference information aggregated from irrelevant regions may disturb the capture of inter-frame consistency relations, leading to wrong matches. To address this issue, we propose Dynamic Cues-Assisted Transformer (DCATr). Firstly, the interference from irrelevant regions is greatly reduced by constraining attention to certain cues, i.e., regions with highly correlated structures of potential corresponding superpoints. Secondly, cues-assisted attention is designed to mine the inter-frame consistency relations, while more attention is assigned to pairs with high consistent confidence in feature aggregation. Finally, a dynamic updating fashion is proposed to facilitate mining richer consistency information, further improving aggregated features' distinctiveness and relieving matching ambiguity. Extensive evaluations on indoor and outdoor standard benchmarks demonstrate that DCATr outperforms all state-of-the-art methods.

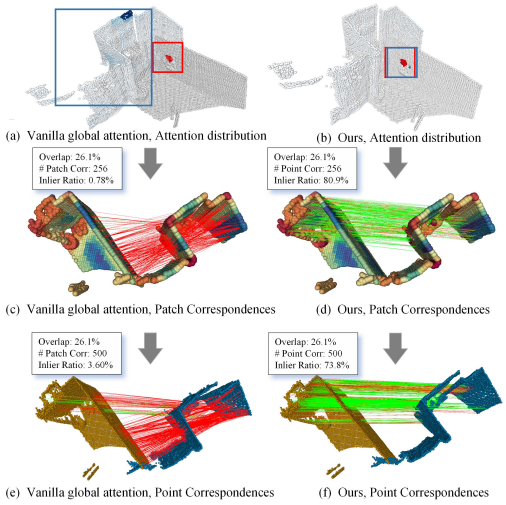

Figure 1. The visualization of the cross-attention heatmaps and matching results. Given two low-overlap point clouds, we sample an anchor patch in one of them and mark it with red. The distribution of cross-attention corresponding to the anchor patch in another point cloud is marked in blue (see the first row). For a clear presentation, they are shown in correctly registered point clouds and highlighted with boxes. The superpoint- and point-matching results are shown in the last two rows, respectively. Green/red lines indicate inliers/outliers.

1. Introduction

Point cloud registration aims to recover the transformation that aligns two partially overlapping point clouds, which is fundamental in numerous computer vision tasks, such as reconstruction [6, 22], pose estimation [27], and simultaneous localization and mapping (SLAM) [4].

Correspondence-based methods [2, 12, 18, 31, 32] have exhibited significant potential and become one of the most popular paradigms in point cloud registration. These methods initially establish point correspondences and derive the relative transformation based on these correspondences. Recent advances are dominated by keypoint-free methods [18, 31, 32] that leverage a coarse-to-fine mechanism to seek correspondences. By partitioning the original point cloud pairs into smaller patches with downsampled superpoints, they first match superpoints and then propagate the results of superpoint matches to individual points, yielding dense correspondences. Consequently, the performance of superpoints matching directly determines the accuracy of point correspondences. Recent methods [18, 31, 32] leverage transformer to learn long-distance dependencies in

*Corresponding author

point clouds to improve the accuracy of superpoint matching. Specifically, these methods utilize the vanilla cross-attention for inter-frame global feature aggregation. Despite the promising progress of these transformer-based methods, they still face the challenge of ambiguous matching. As shown in Fig. 1(c) and Fig. 1(e), in certain scenarios with numerous similar regions and low overlap ratio, the matching accuracy of superpoints (Inlier Ratio) is only 0.78%, resulting in a failure of alignment.

Correct matches tend to satisfy consistency constraints, *i.e.*, they are located in spatially structurally consistent regions, which provides important guidance for capturing the correct inter-frame consistency relations to facilitate point matching [3, 18]. However, due to the lack of consistency constraints, the aggregated features obtained by vanilla global cross-attention inevitably introduce interference from irrelevant regions that do not meet the constraints, resulting in wrong matches. As shown in Fig. 1(a), for the red anchor patch in one point cloud, vanilla global attention (marked with green) distributed in another point cloud is dispersed to irrelevant regions, which means the erroneous information will be aggregated, leading to poor performance of superpoint and point matching (See Fig. 1(c) and Fig. 1(e)). Although previous methods [3, 5, 8, 34] introduce spatial consistency constraints to remove outliers, they only consider information from sparse point pairs, ignoring the richer spatial structure information in entire point clouds. Moreover, they need pre-prepared correspondences as input.

To this end, we propose a Dynamic Cues-Assisted Transformer (DCATr). By explicitly modeling consistency constraints, attention is restricted to regions with potentially consistent structures to capture inter-frame consistency relations, while information from regions with low consistent confidence is suppressed during feature aggregation. Specifically, it consists of the following steps. (1) The potential corresponding superpoints and their cues are extracted, where the cues are those highly geometrically correlated regions of each superpoints. The cues are introduced to provide richer structural information for each superpoint. (2) A cues-assisted attention is designed to mine the inter-frame consistency relations between all superpoint pairs using their cues. Benefiting from the attention is restricted to specific cue regions, the interference of information from irrelevant regions is significantly reduced, and thus, more accurate coherence relations can be captured. Meanwhile, high consistent confidence point pairs receive more attention in feature aggregation (See Fig. 1(b)). (3) To aggregate richer consistency information in point clouds, the point pairs with their cues are dynamically updated several times at the global level. Each time the model finds more possible correspondences or other more discriminative correlated geometric structures, the distinctiveness of aggregated features is improved, ultimately resulting in robust and accurate matching (See Fig. 1(d) and Fig. 1(f)).

Our extensive experiments on indoor and outdoor benchmarks show that DCATr performs favorably against state-of-the-art methods. In summary, our contributions are summarized as follows:

- A novel point cloud matching method is proposed to relieve the matching ambiguity by explicitly modeling inter-frame consistency.
- A cues-assisted attention is proposed to learn the inter-frame consistency relations, which effectively relieves the interference of irrelevant and inconsistent regions in feature aggregation.
- A dynamic update fashion is designed to mine richer and more discriminative information in point clouds, improving aggregated features' distinctiveness.

2. Related Work

Correspondence-based Point Cloud Registration. Early correspondence-based approaches [2, 12] focus on detecting reliable keypoints with descriptors and searching for correspondences based on the detected keypoints. The inherent sparsity of keypoints poses a challenge to repeatability, *i.e.*, sub-sampling introduces a heightened risk that a certain point loses its corresponding point in another frame, limiting the effectiveness of keypoint-based methods. Therefore, more recent and popular keypoint-free approaches use a coarse-to-fine matching mechanism to consider all possible correspondences in point clouds. After establishing correspondences, the robust pose estimators such as RANSAC [9] or other RANSAC-free estimators [3, 5, 8, 18] are used to recover the transformation based on the established correspondences. Our approach inherits the keypoint-free methods, especially enhancing the accuracy of coarse-level matching.

Consistency Modeling. Spatial consistency information serves as a robust guideline for point cloud registration. Current methods [3, 5, 8, 34] primarily utilize this information to identify and remove outliers, typically by establishing various consistency constraints. However, these methods require pre-prepared correspondences. As a prerequisite for these methods, our approach explicitly models inter-frame consistency to facilitate robust and accurate point matching.

Transformers in Point Cloud Registration. Transformers capture the correlation or importance between inputs based on the attention mechanism and are becoming increasingly popular in point cloud registration tasks in recent years [12, 15, 18, 31, 32]. Besides, some methods [18, 32] incorporate the geometric information into attention to encode intra-frame rotation invariance information. Unlike the existing attention mechanism, we propose cues-assisted

attention to mine and aggregate the inter-frame consistency information.

3. Method

Given two partially overlapping point clouds $\mathbf{P} \in \mathbb{R}^{n\times 3}$ and $\mathbf{Q} \in \mathbb{R}^{m\times 3}$, where n and m are the number of points. The goal of our method is to recover the unknown rigid transformation $\mathbf{T} \in SE(3)$ with a rotation matrix $\mathbf{R} \in SO(3)$ and a translation vector $\mathbf{t} \in \mathbb{R}^3$.

An overview of our method is shown in Fig. 2. Our method adopts the coarse-to-fine matching paradigm [18, 31] to extract correspondences. We use KPConv-FPN [25] as our feature backbone, which consists of an encoder-decoder architecture to extract multi-level features. Specifically, given the point cloud $\mathbf{P}$ and $\mathbf{Q}$, the encoder outputs the downsampled coarse-level superpoints $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$ with the associated learned features $\hat{\mathbf{F}}_\mathbf{P} \in \mathbb{R}^{\hat{n}\times\hat{d}}$ and $\hat{\mathbf{F}}_\mathbf{Q} \in \mathbb{R}^{\hat{n}\times\hat{d}}$, respectively. The decoder outputs the fine-level points $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{Q}}$ with associated features $\tilde{\mathbf{F}}_\mathbf{P} \in \mathbb{R}^{\tilde{n}\times\tilde{d}}$ and $\tilde{\mathbf{F}}_\mathbf{Q} \in \mathbb{R}^{\tilde{n}\times\tilde{d}}$, respectively. Here, $\hat{d}$ and $\tilde{d}$ represents the feature dimension. Using the point-to-node grouping strategy [31], the local neighborhood points of each coarse-level superpoint are gathered into a patch. The superpoints $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$ with their features $\hat{\mathbf{F}}_\mathbf{P}$ and $\hat{\mathbf{F}}_\mathbf{Q}$ are fed into the dynamic cues-assisted transformer for feature aggregation (Sec. 3.1) to generate highly representative features for reliable superpoint matching (Sec. 3.2). The results of superpoint matching $\hat{\mathcal{C}} = \left\{(\hat{\mathbf{p}}_i, \hat{\mathbf{q}}_j)|\hat{\mathbf{p}}_i \in \hat{\mathbf{P}}, \hat{\mathbf{q}}_j \in \hat{\mathbf{Q}}\right\}$ are then propagated to fine-level points (Sec. 3.2), yielding dense point correspondences $\tilde{\mathcal{C}} = \left\{(\tilde{\mathbf{p}}_i, \tilde{\mathbf{q}}_j)|\tilde{\mathbf{p}}_i \in \tilde{\mathbf{P}}, \tilde{\mathbf{q}}_j \in \tilde{\mathbf{Q}}\right\}$. The pose estimators such as RANSAC [9] and LGR [18] are introduced to calculate the rotation matrix $\mathbf{R} \in SO(3)$ and the translation vector $\mathbf{t} \in \mathbb{R}^3$ for aligning two point clouds.

3.1. Dynamic Cues-Assisted Transformer

Due to the correct point matches tend to be distributed in inter-frame consistent spatial structures, our goal is to guide the model to focus more on the consistent regions and aggregate the consistency information for obtaining distinctive superpoint features by explicitly modeling inter-frame consistency.

3.1.1 Initial Context Aggregation

For the sake of simplicity, we only introduce $\hat{\mathbf{P}}$ in the following, and the same steps also work for the point cloud $\hat{\mathbf{Q}}$, unless otherwise specified.

Given the point cloud $\hat{\mathbf{P}}$ with their features $\hat{\mathbf{F}}_\mathbf{P} \in \mathbb{R}^{\hat{n}\times\hat{d}}$, a linear layer is first introduced to map $\hat{\mathbf{F}}_\mathbf{P} \in \mathbb{R}^{\hat{n}\times\hat{d}}$ to $\hat{\mathbf{F}}_\mathbf{P} \in \mathbb{R}^{\hat{n}\times d}$. To extract the initial contextual information, we first utilize the geometric self-attention proposed in GeoTrans [18] to mine the intra-frame global geometric features $\hat{\mathbf{G}}_\mathbf{P}^{l=0} \in \mathbb{R}^{\hat{n}\times d}$:

$$\hat{\mathbf{G}}_\mathbf{P}^{l=0} = \mathbf{A}_\mathbf{P}^{s_{l=0}}(\hat{\mathbf{F}}_\mathbf{P}^{l=0}\mathbf{W}^V), \tag{1}$$

where the weight matrix $\mathbf{A}_\mathbf{P}^{s_{l=0}}$ is computed by a row-wise softmax on the attention matrix $\mathbf{E}_\mathbf{P}^{s_{l=0}}$, and $\mathbf{E}_\mathbf{P}^{s_{l=0}}$ is computed as:

$$\mathbf{E}_\mathbf{P}^{s_{l=0}} = \frac{(\hat{\mathbf{F}}_\mathbf{P}^{l=0}\mathbf{W}^Q)(\hat{\mathbf{F}}_\mathbf{P}^{l=0}\mathbf{W}^K + \mathbf{R}\mathbf{W}^R)^T}{\sqrt{d}}. \tag{2}$$

Here, $\mathbf{R} \in \mathbb{R}^{\hat{n}\times d}$ is a geometric structure embedding which consists of a pair-wise distance embedding and a triplet wise angular embedding, $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V, \mathbf{W}^R \in \mathbb{R}^{d\times d}$ are projection matrices for $Query$, Key, $Value$ and geometric embeddings, respectively.

We then apply vanilla cross-attention [12, 18, 29, 31] to exchange information between two point clouds. Given the features $\hat{\mathbf{G}}_\mathbf{P}^{l=0}$ and $\hat{\mathbf{G}}_\mathbf{Q}^{l=0}$ for $\hat{\mathbf{P}}$, $\hat{\mathbf{Q}}$, respectively, the cross-attention features $\hat{\mathbf{Y}}_\mathbf{P}^{l=0} \in \mathbb{R}^{\hat{n}\times d}$ is computed with the $\hat{\mathbf{G}}_\mathbf{Q}^{l=0}$:

$$\hat{\mathbf{Y}}_\mathbf{P}^{l=0} = \mathbf{A}_\mathbf{P}^{c_{l=0}}(\hat{\mathbf{G}}_\mathbf{Q}^{l=0}\mathbf{W}^V), \tag{3}$$

similarly, $\mathbf{A}_\mathbf{P}^{c_{l=0}}$ is computed by a row-wise softmax on the cross-attention score $\mathbf{E}_\mathbf{P}^{c_{l=0}}$, and $\mathbf{E}_\mathbf{P}^{c_{l=0}}$is computed as the feature correlation between the $\hat{\mathbf{G}}_\mathbf{P}^{l=0}$ and $\hat{\mathbf{G}}_\mathbf{Q}^{l=0}$:

$$\mathbf{E}_\mathbf{P}^{c_{l=0}} = \frac{(\hat{\mathbf{G}}_\mathbf{P}^{l=0}\mathbf{W}^Q)(\hat{\mathbf{G}}_\mathbf{Q}^{l=0}\mathbf{W}^K)^T}{\sqrt{d}}. \tag{4}$$

The same cross-attention is also applied in the reverse direction, yielding $\hat{\mathbf{Y}}_\mathbf{Q}^{l=0}$, so that the contextual information aggregates in both directions, $\hat{\mathbf{P}} \rightarrow \hat{\mathbf{Q}}$ and $\hat{\mathbf{Q}} \rightarrow \hat{\mathbf{P}}$.

3.1.2 Consistency Prior Extraction

This module is an extension of the initial context aggregation module, which has a similar structure that mainly contains geometric self-attention and vanilla cross-attention. The big difference is that this module serves for our dynamic updating fashion.

Specifically, given the initial contextual information $\hat{\mathbf{Y}}_\mathbf{P}^{l=0}$ ($\hat{\mathbf{Y}}_\mathbf{Q}^{l=0}$), this module learns the deeper prior $\hat{\mathbf{Y}}_\mathbf{P}^{l=1}$ ($\hat{\mathbf{Y}}_\mathbf{Q}^{l=1}$) by repeating L times. During each repetition, this module consumes the result of the last time, i.e., $\hat{\mathbf{Y}}_\mathbf{P}^{l=j}$ ($\hat{\mathbf{Y}}_\mathbf{Q}^{l=j}$), and output the result of this time $\hat{\mathbf{Y}}_\mathbf{P}^{l=j+1}$ ($\hat{\mathbf{Y}}_\mathbf{Q}^{l=j+1}$), where $j = 1, 2, ..., L$. The learned geometric self-attention matrix in this module is denoted as $\hat{\mathbf{E}}_\mathbf{P}^{s_{l=j}}$ ($\hat{\mathbf{Y}}_\mathbf{Q}^{s_{l=j}}$).

On the one hand, we utilize the learned deeper inter-frame contextual features at each layer to facilitate the search for more possible correspondences and cues in the

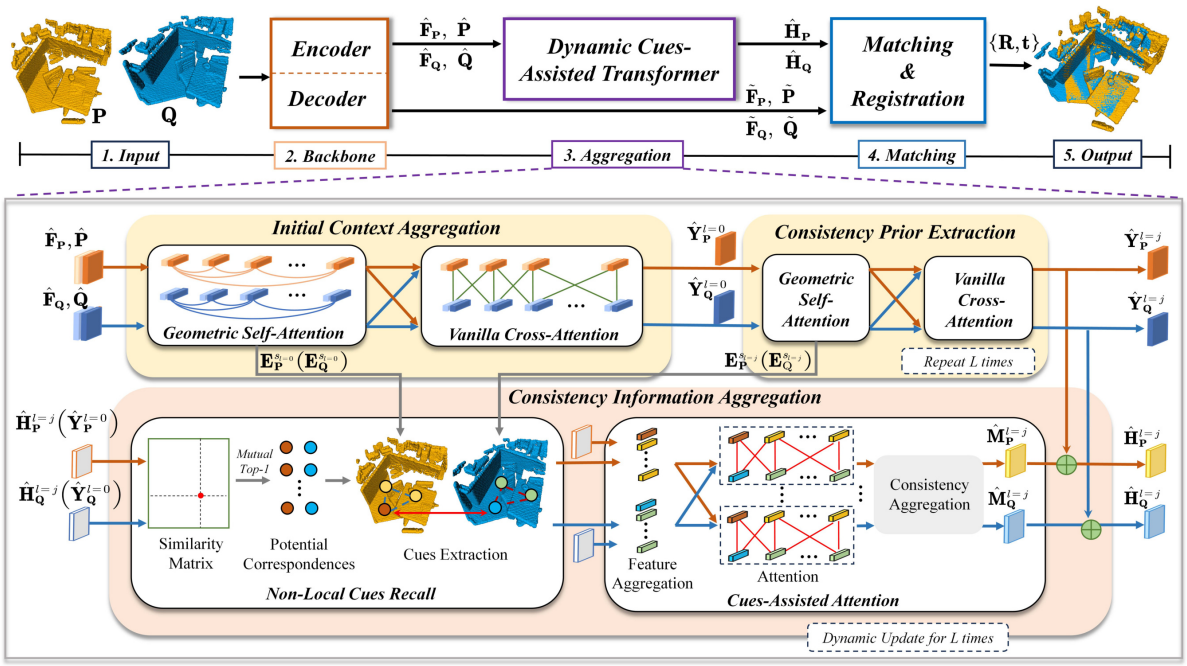

Figure 2. The architecture of our method. The backbone first downsamples the input point clouds and extracts multi-level features. The dynamic cues-assisted transformer (DCATr) enables feature aggregation of coarse-level superpoints (Sec. 3.1). In DCATr, the initial context aggregation module learns the inter-frame geometric features and exchanges information between two point clouds, yielding initial contextual features $\hat{\mathbf{Y}}_{\mathbf{P}}^{l=0}$ ($\hat{\mathbf{Y}}_{\mathbf{Q}}^{l=0}$) (Sec. 3.1.1). The consistency prior extraction module consumes $\hat{\mathbf{Y}}_{\mathbf{P}}^{l=0}$ ($\hat{\mathbf{Y}}_{\mathbf{Q}}^{l=0}$) and mines deeper contextual features $\hat{\mathbf{Y}}_{\mathbf{P}}^{l=j}$ ($\hat{\mathbf{Y}}_{\mathbf{Q}}^{l=j}$) (Sec. 3.1.2) by repeating L times. Meanwhile, the consistency information aggregation module consumes $\hat{\mathbf{Y}}_{\mathbf{P}}^{l=0}$ ($\hat{\mathbf{Y}}_{\mathbf{Q}}^{l=0}$) and the geometric self-attention matrix $\hat{\mathbf{E}}_{\mathbf{P}}^{s_{l=0}}$ ($\hat{\mathbf{E}}_{\mathbf{Q}}^{s_{l=0}}$) to learn the inter-frame consistency relations, yielding the consistency message $\hat{\mathbf{M}}_{\mathbf{P}}^{l=j}$ ($\hat{\mathbf{M}}_{\mathbf{Q}}^{l=j}$) (Sec. 3.1.3), which is implemented in a dynamic updating fashion with L times. The matching and registration module utilizes the coarse-to-fine mechanism to establish correspondences and a pose estimator to calculate the transformation aligning two input point clouds (Sec. 3.2).

global scope. On the other hand, we use the learned geometric self-attention matrix at each layer to capture multi-level inter-frame geometric structure correlations. We refer to the information involved in the above uniformly as the prior information for consistency modeling Sec. 3.1.3.

3.1.3 Consistency Information Aggregation

To model the inter-frame consistency, the non-local cues are first recalled for providing richer spatial structure information. Then, cues-assisted attention is utilized mines and aggregates consistency information. The overall workflow is shown in Fig. 3.

Dynamic Updating. To aggregate richer consistency information in point clouds, we design a dynamic updating fasion. For clarity, let us start with the data flow from the beginning (see Fig. 2). Here only demonstrate for $\hat{\mathbf{P}}$ and the same process acts on $\hat{\mathbf{Q}}$.

Specifically, in the first phase, the consistency information aggregation module (Sec. 3.1.3) consumes the initial contextual features $\hat{\mathbf{Y}}_{\mathbf{P}}^{l=0}$ ($\hat{\mathbf{Y}}_{\mathbf{Q}}^{l=0}$) and the geometric self-attention matrix $\mathbf{E}_{\mathbf{P}}^{s_{l=0}}(\mathbf{E}_{\mathbf{Q}}^{s_{l=0}})$, and outputs the consistency features $\hat{\mathbf{M}}_{\mathbf{P}}^{l=1}$. The consistency prior extraction module consumes the initial contextual features $\hat{\mathbf{Y}}_{\mathbf{P}}^{l=0}$ ($\hat{\mathbf{Y}}_{\mathbf{Q}}^{l=0}$) and outputs the prior $\hat{\mathbf{Y}}_{\mathbf{P}}^{l=1}$ ($\hat{\mathbf{Y}}_{\mathbf{Q}}^{l=1}$)

Start with $l = 1$, in the following phase, the consistency prior extraction module (Sec. 3.1.2) consumes the previous prior $\hat{\mathbf{Y}}_{\mathbf{P}}^{l=j}$ and generate deeper global prior $\hat{\mathbf{Y}}_{\mathbf{P}}^{l=j+1}$, respectively, where $j = 1, 2, ..., L$. At the same time, the global prior $\hat{\mathbf{Y}}_{\mathbf{P}}^{l=j}$ and the consistency features $\hat{\mathbf{M}}_{\mathbf{P}}^{l=j}$ are fused, yielding the hybrid features $\hat{\mathbf{H}}_{\mathbf{P}}^{l=j}$:

$$\hat{\mathbf{H}}_{\mathbf{P}}^{l=j} = \hat{\mathbf{Y}}_{\mathbf{P}}^{l=j} + \hat{\mathbf{M}}_{\mathbf{P}}^{l=j}, \tag{5}$$

and $\hat{\mathbf{H}}_{\mathbf{P}}^{l=j}$ are sent into the consistency information aggregation module (Sec. 3.1.3) in the next times, yielding $\hat{\mathbf{M}}_{\mathbf{P}}^{l=j+1}$.

For simplicity, the symbols in this module are no longer distinguished by superscripts, and we use the $\hat{\mathbf{H}}_{\mathbf{P}}$ ($\hat{\mathbf{H}}_{\mathbf{Q}}$) and $\mathbf{E}_{\mathbf{P}}^{s}$ ($\mathbf{E}_{\mathbf{Q}}^{s}$) represent the input of this module uniformly.

Non-local Cues Recall. The potential correspondences at the global level are first established (see the first step in Fig. 3). Given the point clouds $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$ with their hy-

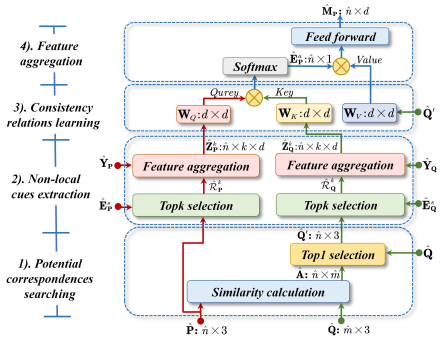

Figure 3. The workflow of the consistency information aggregation module. It includes 4 steps from bottom to top.

brid features $\hat{\mathbf{H}}_\mathbf{P}$ and $\hat{\mathbf{H}}_\mathbf{Q}$, respectively, the similarity matrix by between $\hat{\mathbf{H}}_\mathbf{P}$ and $\hat{\mathbf{H}}_\mathbf{Q}$ is first calculated by $S(i,j) = \tau \left\langle \hat{\mathbf{H}}_\mathbf{P}, \hat{\mathbf{H}}_\mathbf{Q} \right\rangle$, where τ is the temperature coefficient and $\langle , \rangle$ is the inner product. Then, the dual-softmax operator is performed on S to calculate the confidence matrix:

$$\hat{\mathbf{A}}(i,j) = \text{softmax}(S(i,\cdot))(i,j) \cdot \text{softmax}(S(\cdot,j))(i,j). \tag{6}$$

For each point $\hat{\mathbf{p}}_i \in \hat{\mathbf{P}}$, we select the point $\hat{\mathbf{q}}'_j \in \hat{\mathbf{Q}}$ with the highest confidence score to construct the potential correspondence set:

$$\hat{\mathcal{C}}_p = \left\{ (\hat{\mathbf{p}}_i, \hat{\mathbf{q}}'_j) | \hat{\mathbf{q}}'_j = \max(\hat{\mathbf{A}}(\hat{\mathbf{p}}_i, \cdot)) \right\}. \tag{7}$$

To expand the spatial structure information for each point pair, we extract non-local cues using the geometric self-attention matrix $\mathbf{E}^s_\mathbf{P}$ (see the second step in Fig. 3). For simplicity, we only demonstrate for $\hat{\mathbf{P}}$, and the same process acts on $\hat{\mathbf{Q}}$. Given each point $\hat{\mathbf{p}}_i$, we search k points in $\hat{\mathbf{P}}$ by select the Top-k entries in $\mathbf{E}^s_\mathbf{P}$:

$$\hat{\mathcal{R}}^k_p = \text{topk}(\mathbf{E}^s_\mathbf{P}(\hat{\mathbf{p}}_i, \cdot)). \tag{8}$$

Due to the guidance of global geometric correlations in geometric self-attention, the points in $\hat{\mathcal{R}}^k_p$ have the most correlated geometric structures to point $\hat{\mathbf{p}}_i$ and provide richer spatial cues for point $\hat{\mathbf{p}}_i$. The points in $\hat{\mathcal{R}}^k_p$ may be distributed anywhere in the point cloud $\hat{\mathbf{P}}$ and are therefore non-local. We further aggregate the contextual features $\hat{\mathbf{H}}_\mathbf{P}$ of points in $\hat{\mathcal{R}}^k_p$, constructing the cue feature matrix $\hat{\mathbf{Z}}^k_\mathbf{P} \in \mathbb{R}^{\hat{n} \times k \times d}$.

Cues-assisted Attention. Using the cue feature matrix $\hat{\mathbf{Z}}^k_\mathbf{P} \in \mathbb{R}^{\hat{n} \times k \times d}$, we employ cues-assisted attention to mine the consistency relations among all point pairs (see the third step in Fig. 3) and further aggregate consistency information (see the fourth step shown in Fig. 3). Specifically, the attention scores are calculated firstly as the feature correlations between $\hat{\mathbf{z}}^k_{\mathbf{P}_i}$ and $\hat{\mathbf{z}}^k_{\mathbf{Q}_i}$ using the learnable matrices $\mathbf{W}^Q \in \mathbb{R}^{d \times d}$ and $\mathbf{W}^K \in \mathbb{R}^{d \times d}$:

$$\hat{e}^a_{\mathbf{P}_i} = \sum_{l=1}^{k} \sum_{m=1}^{k} \frac{(\hat{\mathbf{z}}^l_{\mathbf{P}_i} \mathbf{W}^Q)(\hat{\mathbf{z}}^m_{\mathbf{Q}_i} \mathbf{W}^K)^T}{\sqrt{d}}. \tag{9}$$

This process learns the consistency relations among all point pairs by means of their cues, and the obtained attention score $\hat{e}^a_{\mathbf{P}_i}$ measures the consistency confidence of each point pair.

The attention score $\hat{e}^a_{\mathbf{P}_i}$ is further used to aggregate the consistency information, followed by a row-wise softmax operator on it. The aggregated features are obtained using the learnable matrices $\mathbf{W}^V \in \mathbb{R}^{d \times d}$ as:

$$\hat{\mathbf{u}}_{\mathbf{P}_i} = \hat{e}^a_{\mathbf{P}_i} (\hat{\mathbf{y}}_{\mathbf{Q}_i} \mathbf{W}^V). \tag{10}$$

Following [26], the obtained feature matrix $\hat{\mathbf{U}}_\mathbf{P}$ constructed by $\hat{\mathbf{u}}_{\mathbf{P}_i}$ is further sent into a feed-forward layer to obtain enhanced message $\hat{\mathbf{M}}_\mathbf{P}$. The cues-assisted cross-attention is also applied in the reverse direction, yielding $\hat{\mathbf{M}}_\mathbf{Q}$.

Analysis. For the cues-assisted attention, on the one hand, the extracted cues expand the spatial structure information of each potential corresponding point, facilitating the consistency relations of point pairs mining. On the other hand, the attention is restricted to cue regions, the interference of irrelevant regional information is significantly reduced, which allows the model to capture the essential inter-frame consistency information more accurately.

For the dynamic updating fasion, after each time the global prior fuse with the consistency features, DCATr re-seeks potential correspondences and re-extracts correlated cues on a global scale. Thus, the point pairs with their cues are dynamic updating, and cues-assisted attention constantly receives different inputs and re-learns consistency relations between them. On the one hand, each time the model finds more reliable correspondences or other more discriminative correlated geometric structures, richer consistency information is aggregated, which significantly improves the distinctiveness of the features. On the other hand, the changing inputs are similar to data augmentation and help cues-assisted attention learn more accurate globally consistent information.

3.2. Matching and Registration Module

We employ a coarse-to-fine paradigm proposed in [31] for feature-based point matching.

Superpoint Matching. In the coarse matching stage, the hybrid features $\hat{\mathbf{H}}^{j=L}_\mathbf{P}$ and $\hat{\mathbf{H}}^{j=L}_\mathbf{Q}$ of the superpoints in the point clouds $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$, respectively, serve as inputs. Following [18], we initially normalize the superpoint features $\hat{\mathbf{H}}^{j=L}_\mathbf{P}$ and $\hat{\mathbf{H}}^{j=L}_\mathbf{Q}$ to a unit hypersphere. For each point's feature $\hat{\mathbf{h}}_{\mathbf{P}_i}$ and $\hat{\mathbf{h}}_{\mathbf{Q}_j}$, we measure pairwise similarities

using a Gaussian correlation matrix $\hat{\mathbf{S}}$, where $\hat{\mathbf{S}}(i,j) = -\exp(-\left\|\hat{\mathbf{h}}_{\mathbf{P}_i} - \hat{\mathbf{h}}_{\mathbf{Q}_j}\right\|_2^2)$. After dual normalization of $\hat{\mathbf{S}}$ to capture global feature correlations [18, 20, 24, 32], we select the Top-$\hat{k}$ entries in $\hat{\mathbf{S}}$ as the coarse correspondence set $\hat{\mathcal{C}} = \left\{(\hat{\mathbf{p}}_i, \hat{\mathbf{q}}_j) | \hat{\mathbf{p}}_i \in \hat{\mathbf{P}}, \hat{\mathbf{q}}_j \in \hat{\mathbf{Q}}\right\}$.

Point Matching. In the fine-level point matching, the point-to-node strategy [18, 31, 32] is utilized to assign dense points $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{Q}}$ to its closest superpoint $\hat{\mathbf{p}}_i \in \hat{\mathbf{P}}$, $\hat{\mathbf{q}}_i \in \hat{\mathbf{Q}}$, respectively. The group of points assigned to each superpoint $\hat{\mathbf{p}}_i$ is denoted as $\tilde{\mathbf{B}}_{\mathbf{P}_i}$, where $\tilde{\mathbf{B}}_{\mathbf{P}_i} \subseteq \tilde{\mathbf{P}}$, and the features associated with $\tilde{\mathbf{B}}_{\mathbf{P}_i}$ is defined as $\tilde{\mathbf{B}}^i_{\mathbf{F_P}}$, where $\tilde{\mathbf{B}}^i_{\mathbf{F_P}} \subseteq \tilde{\mathbf{F}}_{\mathbf{P}}$. The similarity between feature groups $\tilde{\mathbf{B}}^i_{\mathbf{F_P}}$ and $\tilde{\mathbf{B}}^j_{\mathbf{F_Q}}$ is calculated as $\tilde{\mathbf{S}}_g = \tilde{\mathbf{B}}^i_{\mathbf{F_P}}(\tilde{\mathbf{B}}^j_{\mathbf{F_Q}})^T/\sqrt{\tilde{d}}$. Subsequently, we follow the approach in [18, 21] to use the Sinkhorn Algorithm [23] to obtain a normalized confidence matrix $\bar{\mathbf{S}}_g$. We select the mutual Top-$\tilde{k}$ entries in $\bar{\mathbf{S}}_g$, *i.e.*, entries with the Top-$\tilde{k}$ confidence values in both the rows and columns, to form a point correspondence $\tilde{C}_g$. The final correspondence set $\tilde{C}$ is constructed as $\tilde{C} = \bigcup_{g=1}^{|\tilde{C}|} \tilde{C}_g$.

Regsitration. After obtaining the dense point correspondences, a pose estimator such as RANSAC [9] and LGR [18]is introduced to calculate the transformation for outputting the registered point clouds.

Loss Function. We use the same superpoint matching loss $\mathcal{L}_c$ and point matching loss $\mathcal{L}_f$ as [18], and the overall loss function is $\mathcal{L} = \mathcal{L}_c + \mathcal{L}_f$.

4. Experiments

We evaluate our method on indoor 3DMatch [33], 3DLoMatch [12] (Sec. 4.1) and outdoor KITTI odometry [10] benchmarks (Sec. 4.2).

Experimental Setup. We conduct our experiments with DCATr using PyTorch [17] on an Intel I7 12700 CPU and an NVIDIA RTX 4090 GPU. The Adam optimizer [14] is utilized to train our model with an initial learning rate of 0.0001 and weight decay of 0.000001. The step learning approach is applied with a decay rate of 0.95. We set the feature dimension $\hat{d}$, $\tilde{d}$, and d to 1024, 256, and 256 for 3DMatch and 3DLoMatch, respectively, while they are set to 2048, 256, and 100 for KITTI odometry, respectively. We set the times of repetition L as 3, and the number of cues k for each potential corresponding point is set to 6.

4.1. Indoor Benchmarks: 3DMatch & 3DLoMatch

Dataset. 3DMatch [33] covers 62 indoor scenes, of which 46 were used for training, 8 for validation, and 8 for testing. We found our evaluation on the pre-processed point clouds provided by [12]. We follow the protocols of 3DMatch and 3DLoMatch [12], where the overlap between point cloud pairs is greater than 30% in 3DMatch and is in $10\% \sim 30\%$ in 3DLoMatch.

#Samples	3DMatch					3DLoMatch				
	250	500	1000	2500	5000	250	500	1000	2500	5000
Inlier Ratio(%) ↑										
3DSNet [11]	16.4	21.5	26.4	32.5	36.0	4.8	6.4	8.0	10.1	11.4
FCGF [7]	34.1	42.5	48.7	54.1	56.8	11.6	14.8	17.2	20.0	21.4
D3Feat [2]	39.0	38.8	40.4	41.5	41.8	13.2	13.1	14.0	14.6	15.0
D3Feat [2]	41.8	41.5	40.4	38.8	39.0	15.0	14.6	14.0	12.1	13.2
SpinNet [1]	27.6	33.9	39.4	44.7	47.5	11.1	13.8	16.3	19.0	20.5
YOHO [28]	41.2	46.4	55.7	60.7	64.4	15.0	18.2	22.6	23.2	25.9
Predator [12]	49.3	54.1	57.1	58.4	58.0	25.8	27.5	28.3	28.1	26.7
CofiNet [31]	52.2	52.2	51.9	51.2	49.8	26.9	26.8	26.7	25.9	24.4
GeoTrans [18]	85.1	82.2	76.0	75.2	71.9	57.7	52.9	46.2	45.3	43.5
DCATr (Ours)	**87.6**	**86.5**	**84.7**	**81.0**	**76.5**	**62.1**	**60.3**	**57.9**	**53.3**	**48.4**
Registration Recall (%) ↑										
3DSNet [11]	50.8	67.6	71.4	76.2	78.4	11.0	17.0	23.3	29.0	33.0
FCGF [7]	71.4	81.6	83.3	84.7	85.1	26.8	35.4	38.2	41.7	40.1
D3Feat [2]	77.9	82.4	83.4	84.5	81.6	39.1	43.8	46.9	42.7	37.2
SpinNet [1]	70.2	883.5	85.5	86.6	88.6	26.8	39.8	48.3	54.9	59.8
YOHO [28]	84.5	88.6	89.1	90.3	90.8	48.0	56.5	63.2	65.5	65.2
Predator [12]	86.6	88.5	90.6	89.9	89.0	58.1	60.8	62.4	61.2	59.8
CofiNet [31]	87.0	87.4	88.4	88.9	89.3	61.0	63.1	64.2	66.2	67.5
GeoTrans [18]	91.2	91.4	91.8	91.8	92.0	73.5	74.1	74.2	74.8	75.0
DCATr (Ours)	**91.6**	**91.9**	**92.2**	**92.4**	**92.6**	**73.7**	**75.1**	**75.7**	**76.4**	**76.8**
Feature Matching Recall (%) ↑										
3DSNet [11]	82.9	90.1	92.9	94.3	95.0	34.2	45.2	53.6	61.7	63.6
FCGF [7]	96.6	96.7	97.0	97.3	97.4	67.3	71.7	74.2	75.4	76.6
D3Feat [2]	93.1	94.1	94.5	95.4	95.6	66.5	66.7	67.0	66.7	67.3
SpinNet [1]	94.3	95.5	96.8	97.2	97.6	63.6	70.0	72.5	74.9	75.3
YOHO [28]	96.0	97.7	97.5	97.6	98.2	69.1	73.8	76.3	78.1	79.4
Predator [12]	96.5	96.3	96.5	96.6	96.6	75.3	75.7	76.3	77.4	78.6
CofiNet [31]	**98.3**	98.2	98.1	**98.3**	**98.1**	82.6	83.1	83.3	83.5	83.1
GeoTrans [18]	97.6	97.9	97.9	97.9	97.9	**88.3**	**88.6**	**88.8**	**88.6**	**88.3**
DCATr (Ours)	98.2	**98.3**	**98.4**	98.0	**98.1**	87.7	87.5	87.7	87.2	87.4

Table 1. Evaluation results on 3DMatch and 3DLoMatch based on RANSAC pose estimator with a varying number of correspondences.

Metrics. We report five metrics: (1) Inlier Ratio (IR), the proportion of putative correspondences whose residuals are below a specific threshold (e.g., 0.1m) under the ground-truth transformation, (2) Registration Recall (RR), the fraction of point cloud pairs for which the transformation error is less than a particular threshold (e.g., RMSE < 0.2m), (3) Feature Matching Recall (FMR), the fraction of point cloud pairs whose inlier ratio is above a certain threshold (e.g., 5%), (4) Relative Rotation Error (RRE), the geodesic distance between estimated and ground-truth rotation matrices, (5) Relative Translation Error (RTE), the Euclidean distance between estimated and ground-truth translation vectors. We compare our method with the recent state of the arts: 3DSNet[11], FCGF [7], D3Feat [2], SpinNet[1], YOHO[28], Predator [12], CoFiNet [31], and GeoTrans [19].

Results with RANSAC Estimator. we run 50K RANSAC [9] iterations and report the results with different numbers of correspondences in Tab. 1. For Inlier ratio, DCATr achieves the highest performance on both datasets, which means the highest accuracy of point matching. In particular, compared with the sota method GeoTrans, DCATr outperforms by a considerable margin in IR, up to $2.5\% \sim 11.7\%$. For Registration Recall, DCATr also achieves the highest performance on both datasets. Note that the Registration Recall reflects the performance of the final reg-

Methods	3DMatch				
	RR(%) ↑	IR(%) ↑	FMR(%) ↑	RRE(°) ↓	RTE(m) ↓
FCGF [7]	83.3	48.7	97.0	1.949	0.066
D3Feat [2]	83.4	40.4	94.5	2.161	0.067
Predator [12]	90.6	57.1	96.5	2.029	0.064
CoFiNet [31]	88.4	51.9	98.1	2.011	2.011
GeoTrans [18]	91.5	70.3	97.7	1.625	0.053
DCATr(Ours)	**92.1**	**75.0**	**98.1**	**1.536**	**0.050**
Methods	3DLoMatch				
	RR(%) ↑	IR(%) ↑	FMR(%)↑	RRE(°) ↓	RTE(m) ↓
FCGF [7]	38.2	17.2	74.2	3.147	0.100
D3Feat [2]	46.9	14.0	67.0	3.361	0.103
Predator [12]	61.2	28.3	76.3	3.048	0.093
CoFiNet [31]	64.2	26.7	83.3	3.280	0.094
GeoTrans [18]	74.0	43.3	**88.1**	2.547	0.074
DCATr(Ours)	**75.7**	**48.2**	87.3	**2.445**	**0.072**

Table 2. Registration results on 3DMatch and 3DLoMatch based on LGR pose estimator.

Methods	3DMatch				3DLoMatch			
	RR(%)	PIR(%)	IR(%)	FMR(%)	RR(%)	PIR(%)	IR(%)	FMR(%)
GeoTrans	91.5	86.1	70.3	97.7	74.0	54.9	43.3	**88.1**
Static (serial)	91.0	86.0	66.8	98.0	72.8	54.2	41.3	87.7
Static (residual)	91.0	86.6	70.9	**98.4**	74.2	55.2	43.9	87.3
Dynamic (residual)	**92.1**	**88.1**	**75.0**	98.1	**75.7**	**58.0**	**48.2**	87.3

Table 3. Ablation studies about cues-assisted attention.

istration, which means DCATr achieves more successfully matched point cloud pairs. For Feature Matching Recall, DCATr achieves comparable performance.

Results with LGR Estimator. We use the LGR estimator proposed in [18]. As shown in Tab. 2, DCATr achieves the highest performance in RR, FMR, RRE, and RTE on 3DMatch, and achieves the highest performance in RR, IR, and RRE on 3DLoMatch, with a comparable performance in other metrics. In particular, DCATr improves by 2.3% in RR and 5.1% in IR on 3DLoMatch.

Qualitative Results of Registration. We visualize a collection of the registration results of GeoTrans [18] and DCATr in Fig. 4. Benefitting from the aggregation of consistency information, in scenarios with low overlap and a large number of geometrically indistinct (1^{st} row) or similar structures (2^{nd} and 3^{rd} rows), DCATr is still able to effectively avoid the interference of similar but non-overlapping regions, obtaining high-performance patch- and point-matching results.

Qualitative Results of Extracted Dynamic Cues. The visualization of dynamic cues are shown in Fig. 5. The red box marks the anchor regions, while regions marked with green and blue represent the cues in source and target point clouds, respectively. The brown boxes indicate when green and blue cues overlap. It shows that DCATr extracts more consistent cues after more updating times.

Ablation Studies. To verify the impact of the dynamical updating fashion on the performance of DCATr, we first implemented two connection structures for the prior information extraction modules and cues-assisted attention modules. One is to connect the cues-assisted attention modules directly after the prior information extraction modules, similar to a serial connection. Another is residual connection that shown in Fig. 2 (i.e., extracting hybrid features). At the same time, we implemented a static fashion. We use the LGR pose estimator in all ablation studies. In Tab. 3, the static fashion using serial connection (2^{nd} row) results are lower than GeoTrans because that the way of serial connection may affect the prior information learning. While the static fashion using residual connection (3^{rd} row) results are slightly better than GeoTrans. Moreover, the consistency information mined by the static fashion is limited, and our dynamic fashion (using residual connection, (4^{th} row)) mines richer cues through dynamic updates and achieves greater performance gains.

Methods	3DMatch			3DLoMatch		
	RR(%)	IR(%)	FMR(%)	RR(%)	IR(%)	FMR((%)
GeoTrans+SC2-PCR	**92.4**	70.9	**98.2**	74.1	43.5	87.1
Ours	92.1	**75.0**	98.1	**75.7**	**48.2**	**87.3**

Table 4. Registration results with SC2-PCR.

Methods	3DMatch (rotated)			3DLoMatch (rotated)		
	RR(%)	IR(%)	FMR(%)	RR(%)	IR(%)	FMR(%)
GeoTrans	92.0	68.2	**97.8**	71.8	**40.0**	**85.8**
Ours	**94.9**	**69.7**	**97.8**	**73.6**	39.6	84.3

Table 5. Generalization results under full-range rotations.

Comparison with the Method Modeling Consistency. Existing methods utilize spatial consistency information to remove outliers, among which SC2-PCR[5] is a typical and convenient method. To verify the effectiveness of DCATr in introducing the idea of consistency modeling into the correspondence extraction session, we feed the geotans-extracted correspondences into SC2-PCR for registration, and compare them directly with DCATr. Tab. 4 shows the superiority of DCATr, especially at low overlap.

Generalization under Full-range Rotations. We also verifies the generalization of DCATr under full-range rotations. Here we use RANSAC-50K estimator. Tab. 5 shows that our method can maintain the performance under strong rotation even though we didn't customize for this situation.

4.2. Outdoor Benchmark: KITTI

Dataset. The KITTI odometry dataset [10] consists of 11 sequences capturing outdoor driving scenarios using LiDAR scans. Following the methods described in [2, 7, 12, 19], we use sequences 0-5 for training, 6-7 for validation, and 8-10 for testing. Meanwhile, we refined the ground-truth poses using the Iterative Closest Point (ICP) algorithm and conducted evaluations only on point cloud pairs that are at least 10 meters apart.

Metrics. We adopted three metrics similar to [12, 18] to

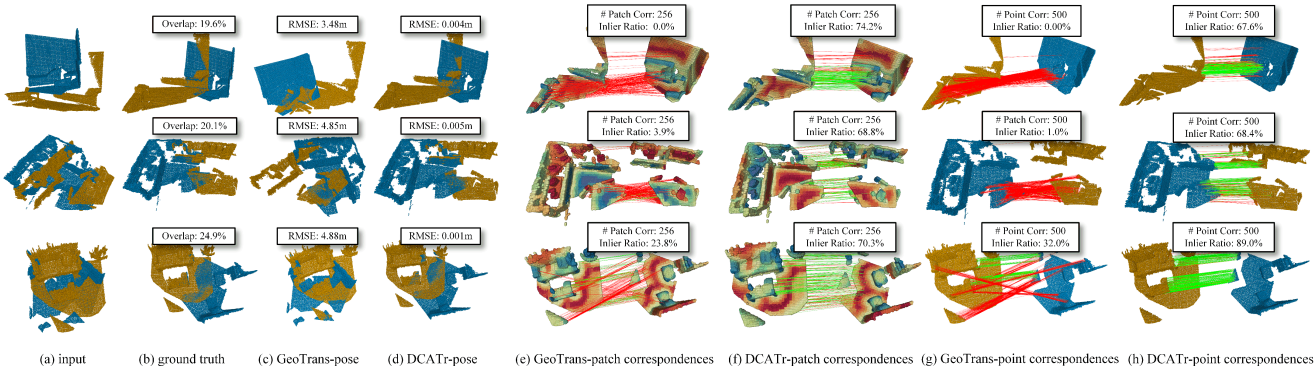

Figure 4. Qualitative results on 3DLoMatch of GeoTrans and DCATr. Each row shows the point cloud in a different scenario, and columns 3 to 8 show a comparison of different metrics. Green/red lines indicate inliers/outliers. Benefitting from the consistency information aggregation, DCATr can capture consistent regions in geometrically insignificant scenes as well as in a large number of similarly structured scenes, which greatly improves the matching accuracy (see (f) and (h)) and the registration performance (see (d)).

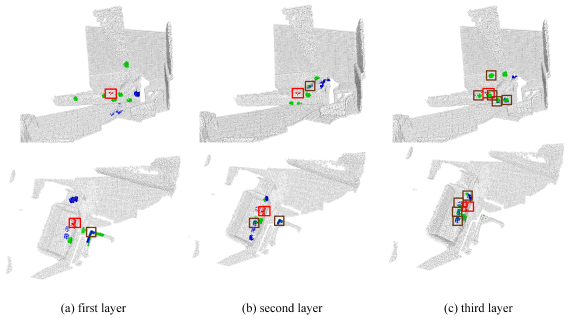

Figure 5. The visualization of dynamic cues.

Methods	RTE(cm) ↓	RRE(°) ↓	RR(%) ↑
3DFeatNet [30]	25.9	0.25	96.0
FCGF [7]	9.5	0.30	96.6
D3Feat [2]	7.2	0.30	99.8
FMR [13]	∼66	1.49	90.6
DGR [8]	∼32	0.37	98.7
SpinNet [1]	9.9	0.47	99.1
Predator [12]	6.8	0.27	99.8
HregNet [16]	∼12	0.29	99.7
CofiNet [31]	8.2	0.41	99.8
GeoTrans [18]	6.8	0.24	**99.8**
DCATr (Ours)	**6.6**	**0.22**	99.7

Table 6. Registration results on KITTI odometry.

assess the performance of our method. These metrics include (1) Relative Rotation Error (RRE), the geodesic distance between estimated and ground-truth rotation matrices, (2) Relative Translation Error (RTE), the Euclidean distance between estimated and ground-truth translation vectors, and (3) Registration Recall (RR), the fraction of point cloud pairs meeting specific threshold criteria (i.e., RRE$< 5°$ and RTE< 2m). We compare our method with the state-of-the-art methods: 3DFeat-Net [30], FCGF [7], D3Feat [2], FMR [13], DGR [8], SpinNet[1], Predator [12], HRegNet [16], CoFiNet [31] and GeoTrans [19].

Registration Results. We use the LGR pose estimator, and the registration results are shown in Tab. 6. Compared to the previous method, our model achieves lower RRE and RTE, and has a comparable performance in RR, showing good generality of DCATr on outdoor scenes.

5. Conclusion

We introduced DCATr, a dynamic cues-assisted transformer for robust point cloud registration. In DCATr, the potential corresponding superpoints and their highly geometric correlated structures (the cues) are extracted. Then the model's attention is restricted to these cue regions, which relieves the distraction from irrelevant regions. Further, we designed cues-assisted attention to learn the consistency relations among the cues, while suppressing the information from inconsistent regions in feature aggregation. Finally, by mining more possible potential corresponding superpoints and their cue regions, a dynamic updating fashion is proposed, which significantly improves the distinctiveness of aggregated features. Extensive evaluations are conducted on indoor and outdoor standard benchmarks to demonstrate the superiority of our DCATr.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grant 62371201, by the Basic Research Surpport Plan of HUST (No.6142113-JCKY2022003), and by the China Scholarship Council for funding visiting Ph.D. student (No. 202106160054).

References

[1] Sheng Ao, Qingyong Hu, Bo Yang, Andrew Markham, and Yulan Guo. Spinnet: Learning a general surface descriptor for 3d point cloud registration. In *CVPR*, pages 11753–11762, 2021. 6, 8

[2] Xuyang Bai, Zixin Luo, Lei Zhou, Hongbo Fu, Long Quan, and Chiew-Lan Tai. D3feat: Joint learning of dense detection and description of 3d local features. In *CVPR*, 2020. 1, 2, 6, 7, 8

[3] Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. Pointdsc: Robust point cloud registration using deep spatial consistency. In *CVPR*, pages 15859–15869, 2021. 2

[4] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE TR-o*, 32(6):1309–1332, 2016. 1

[5] Zhi Chen, Kun Sun, Fan Yang, and Wenbing Tao. Sc2-pcr: A second order spatial compatibility for efficient and robust point cloud registration. In *CVPR*, pages 13221–13231, 2022. 2, 7

[6] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *CVPR*, pages 5556–5565, 2015. 1

[7] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *ICCV*, pages 8958–8966, 2019. 6, 7, 8

[8] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *CVPR*, pages 2514–2523, 2020. 2, 8

[9] Martin A Fishler. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6): 381–395, 1981. 2, 3, 6

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 6, 7

[11] Zan Gojcic, Caifa Zhou, Jan D Wegner, and Andreas Wieser. The perfect match: 3d point cloud matching with smoothed densities. In *CVPR*, pages 5545–5554, 2019. 6

[12] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *CVPR*, 2021. 1, 2, 3, 6, 7, 8

[13] Xiaoshui Huang, Guofeng Mei, and Jian Zhang. Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences. In *CVPR*, pages 11366–11374, 2020. 8

[14] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Comput. Sci.*, 2014. 6

[15] Yang Li and Tatsuya Harada. Lepard: Learning partial point cloud matching in rigid and deformable scenes. In *CVPR*, pages 5554–5564, 2022. 2

[16] Fan Lu, Guang Chen, Yinlong Liu, Lijun Zhang, Sanqing Qu, Shu Liu, and Rongqi Gu. Hregnet: A hierarchical network for large-scale outdoor lidar point cloud registration. In *ICCV*, pages 16014–16023, 2021. 8

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NIPS*, 32, 2019. 6

[18] Zheng Qin, Hao Yu, Changiian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *CVPR*, pages 11133–11142, 2022. 1, 2, 3, 5, 6, 7, 8

[19] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, Slobodan Ilic, Dewen Hu, and Kai Xu. Geotransformer: Fast and robust point cloud registration with geometric transformer. *IEEE TPAMI*, 45(8):9806–9821, 2023. 6, 7, 8

[20] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *NIPS*, 31, 2018. 6

[21] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, pages 4938–4947, 2020. 6

[22] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 1

[23] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. 6

[24] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, pages 8922–8931, 2021. 6

[25] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, Francois Goulette, and Leonidas Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *ICCV*, 2019. 3

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 30, 2017. 5

[27] Richard Vock, Alexander Dieckmann, Sebastian Ochmann, and Reinhard Klein. Fast template matching and pose estimation in 3d point clouds. *Comput Graph*, 79:36–45, 2019. 1

[28] Haiping Wang, Yuan Liu, Zhen Dong, and Wenping Wang. You only hypothesize once: Point cloud registration with rotation-equivariant descriptors. In *ACM MM*, pages 1630–1641, 2022. 6

[29] Yue Wang and Justin M Solomon. Deep closest point: Learning representations for point cloud registration. In *ICCV*, pages 3523–3532, 2019. 3

[30] Zi Jian Yew and Gim Hee Lee. 3dfeat-net: Weakly supervised local 3d features for point cloud registration. In *ECCV*, pages 607–623, 2018. 8

[31] Hao Yu, Fu Li, Mahdi Saleh, Benjamin Busam, and Slobodan Ilic. Cofinet: Reliable coarse-to-fine correspondences for robust point cloud registration. *NIPS*, 34:23872–23884, 2021. 1, 2, 3, 5, 6, 7, 8

[32] Hao Yu, Zheng Qin, Ji Hou, Mahdi Saleh, Dongsheng Li, Benjamin Busam, and Slobodan Ilic. Rotation-invariant

transformer for point cloud matching. In *CVPR*, pages 5384–5393, 2023. 1, 2, 6

[33] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *CVPR*, pages 1802–1811, 2017. 6

[34] Xiyu Zhang, Jiaqi Yang, Shikun Zhang, and Yanning Zhang. 3d registration with maximal cliques. In *CVPR*, pages 17745–17754, 2023. 2