

REGTR: End-to-end Point Cloud Correspondences with Transformers

Zi Jian Yew Gim Hee Lee
National University of Singapore

yewzijian@u.nus.edu gimhee.lee@comp.nus.edu.sg

Abstract

Despite recent success in incorporating learning into point cloud registration, many works focus on learning feature descriptors and continue to rely on nearest-neighbor feature matching and outlier filtering through RANSAC to obtain the final set of correspondences for pose estimation. In this work, we conjecture that attention mechanisms can replace the role of explicit feature matching and RANSAC, and thus propose an end-to-end framework to directly predict the final set of correspondences. We use a network architecture consisting primarily of transformer layers containing self and cross attentions, and train it to predict the probability each point lies in the overlapping region and its corresponding position in the other point cloud. The required rigid transformation can then be estimated directly from the predicted correspondences without further post-processing. Despite its simplicity, our approach achieves state-of-the-art performance on 3DMatch and ModelNet benchmarks. Our source code can be found at <https://github.com/yewzijian/RegTR>.

1. Introduction

Rigid point cloud registration refers to the problem of finding the optimal rotation and translation parameters that align two point clouds. A common solution to point cloud registration follows the following pipeline: 1) detect salient keypoints, 2) compute feature descriptors for these keypoints, 3) obtain putative correspondences via nearest neighbor matching, and 4) estimate the rigid transformation, typically in a robust fashion using RANSAC. In recent years, researchers have applied learning to point cloud registration. Many of these works focus on learning the feature descriptors [14, 17, 61] and sometimes also the keypoint detection [2, 23, 56]. The final two steps generally remain unchanged and these approaches still require nearest neighbor matching and RANSAC to obtain the final transformation. These algorithms do not take the post-processing into account during training, and their performance can be sensitive to the post-processing choices to pick out the correct

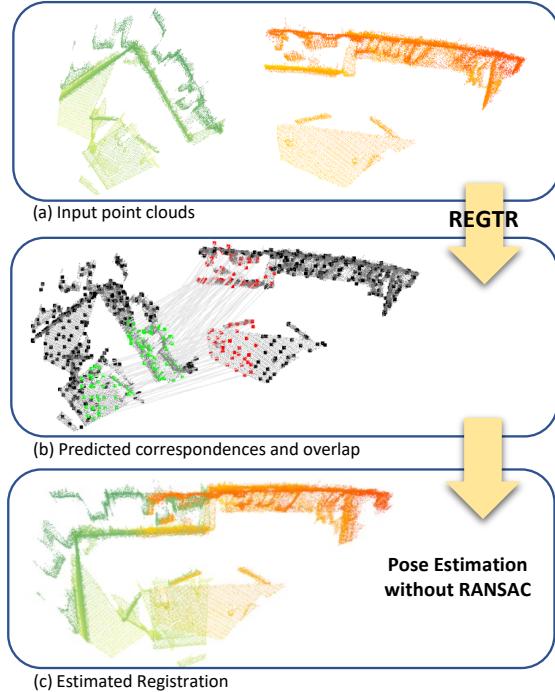


Figure 1. Our network directly outputs final correspondences and the overlap scores. The required rigid transformation can then be directly computed from these correspondences without RANSAC.

correspondences, e.g. number of sampled interest points or distance threshold in RANSAC.

Several works [50, 51, 57] avoid the non-differentiable nearest neighbor matching and RANSAC steps by estimating the alignment using soft correspondences computed from the local feature similarity scores. In this work, we take a slightly different approach. We observe that the learned local features in these works are mainly used to establish correspondences. Thus, we focus on having the network directly predict a set of clean correspondences instead of learning good features. We are motivated by the recent line of works [7, 34] which make use of transformer attention [49] layers to predict the final outputs for various tasks with minimal post-processing. Although attention mechanisms have previously been used in registration of both point clouds [23, 50] and images [39], these works utilize

attention layers mainly to aggregate contextual information to learn more discriminative feature descriptors. A subsequent RANSAC or optimal transport step is still often used to obtain the final correspondences. In contrast, our Registration Transformer (REGTR) utilizes attention layers to directly output a consistent set of *final* point correspondences, as illustrated in Fig. 1. Since our network outputs clean correspondences, the required rigid transformation can be estimated directly without additional nearest neighbor matching and RANSAC steps.

Our REGTR first uses a point convolutional backbone [45] to extract a set of features while downsampling the input pair of point clouds. The features of both point clouds are passed into several transformer [49] layers consisting of multi-head self and cross attentions to allow for global information aggregation, while taking into account the point positions through positional encodings to allow the network to utilize rigidity constraints to correct bad correspondences. The resulting features are then used to predict the corresponding transformed locations of the downsampled points. We additionally predict overlap probability scores to weigh the predicted correspondences when computing the rigid transformation. Unlike the more common approach of computing correspondences via nearest neighbor feature matching, which requires interest points to be present at the same locations in both point clouds, our network is trained to directly predict corresponding point locations. As a result, we do not require sampling large number of interest points (*e.g.* in [23, 61]) or a keypoint detector (*e.g.* [30, 62]) that produces repeatable points. Instead, we establish correspondences on simple grid subsampled points.

Although our REGTR is simple in design, it achieves state-of-the-art performance on the 3DMatch [61] and ModelNet [52] datasets. It also has fast run times since it does not require running RANSAC on a large number of putative correspondences. In summary, our contributions are:

- We directly predict a consistent set of final point correspondences via self and cross attention, without using the commonly used RANSAC nor optimal transport layers.
- We evaluate on several datasets and demonstrate state-of-the-art performance, achieving precise alignments despite using a small number of correspondences.

2. Related Work

Correspondence-based registration. Correspondence-based approaches for point cloud registration first establish correspondences between salient keypoints, followed by robust estimation of the rigid transformation. To accomplish the first step, many keypoint detectors [43, 62] and feature descriptors [38, 46] have been handcrafted. Pioneered by 3DMatch [61], many researchers propose to improve feature descriptors [14, 17, 27, 61] and also keypoint

detection [2, 30, 56] by learning from data. Recently, Predator [23] utilizes attention mechanisms to aggregate contextual information to learn more discriminative feature descriptors. Most of these works are trained by optimizing a variant of the contrastive loss [11, 41] between feature descriptors of matching and non-matching points, and rely on a subsequent nearest neighbor matching step and RANSAC to select the correct correspondences.

Learned direct registration methods. Instead of combining learned descriptors with robust pose estimation, some works incorporate the entire pose estimation into the training pipeline. Deep Closest Point (DCP) [50] proposes a learned version of Iterative Closest Point (ICP) [4, 8], and utilizes soft correspondences on learned pointwise features to compute the rigid transform in a differentiable manner. However, DCP cannot handle partial overlapping point clouds and thus later works overcome the limitation by detecting keypoints [51] or using optimal transport layers with an added slack row and column [19, 57]. IDAM [31] considers both feature and Euclidean space during its pairwise matching process. PCAM [6] multiplies the cross-attention matrices at multiple levels to fuse low and high-level contextual information. DeepGMR [60] learns to compute point-to-distribution correspondences. A separate group of works [1, 24, 32, 40, 55] rely on global feature descriptors and circumvent the local feature correspondence step. PointNetLK [1] aligns two point clouds by minimizing the distances between their global PointNet [37] features, in a procedure similar to the Lucas-Kanade [3] algorithm. Li *et al.* [32] extends it to use analytical Jacobians to improve the generalization behavior. OMNet [55] incorporates masking into the global feature to better handle partial overlapping point clouds. Our method utilizes local features and is similar to *e.g.* [50, 51], but we focus on predicting accurate corresponding point locations via transformer attention layers.

Learned correspondence filtering. The putative correspondences obtained from correspondence-based methods contain outliers, and thus RANSAC is typically used to filter out wrong matches when estimating the required transformation. However, RANSAC is non-differentiable and cannot be used within a training pipeline. Recent works alleviate this problem by modifying RANSAC to enforce differentiability [5], or by learning to identify which of the putative correspondences are inliers [13, 20, 29, 58]. In addition to inlier classification, 3DRegNet [36] also regresses the rigid transformation parameters using a deep network. Different from these works, we directly predict the clean correspondences without explicitly computing the noisy putative correspondences.

Transformers. Transformers [49] propose a novel attention mechanism that makes use of multiple layers of self and cross multi-head attention to exchange information between the input and output. Although originally designed for NLP

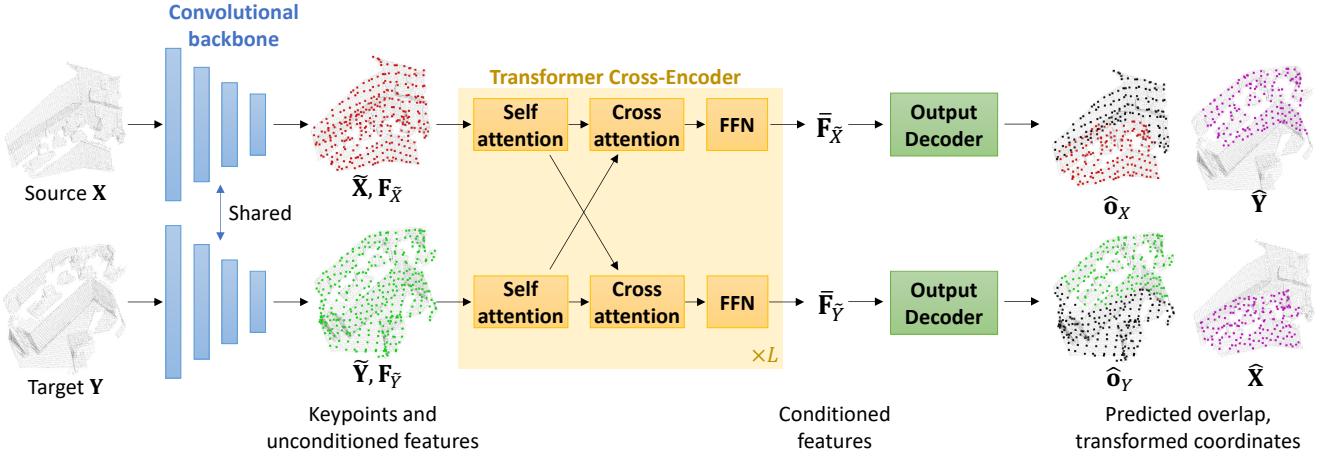


Figure 2. REGTR uses the KPConv convolutional backbone to extract a set of features for a sparse set of points. The features are then passed into several transformer cross-encoder layers. Lastly, the output decoder predicts the overlap score and the corresponding transformed coordinates of the sparse keypoints, which can be used for direct estimation of the pose. Best viewed in color.

tasks, the attention mechanism has recently been shown to be useful for many computer vision tasks [7, 18, 34, 59], and we utilize it in our work to predict point correspondences.

3. Problem Definition

Consider two point clouds $\mathbf{X} \in \mathbb{R}^{M \times 3}$ and $\mathbf{Y} \in \mathbb{R}^{N \times 3}$, which we denote as the *source* and *target*, respectively. The objective of point cloud registration is to recover the unknown rigid transformation consisting of a rotation $\mathbf{R} \in \text{SO}(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$ that aligns \mathbf{X} to \mathbf{Y} .

4. Our Approach

Figure 2 illustrates our overall framework. We first convert the input point clouds into a smaller set of downsampled keypoints $\tilde{\mathbf{X}} \in \mathbb{R}^{M' \times 3}$ and $\tilde{\mathbf{Y}} \in \mathbb{R}^{N' \times 3}$ with $M' < M, N' < N$, and their associated features $\mathbf{F}_{\tilde{\mathbf{X}}} \in \mathbb{R}^{M' \times D}, \mathbf{F}_{\tilde{\mathbf{Y}}} \in \mathbb{R}^{N' \times D}$ (Sec. 4.1). Our network then passes these keypoints and features into several transformer cross-encoder layers (Sec. 4.2) before finally outputting the corresponding transformed locations $\hat{\mathbf{Y}} \in \mathbb{R}^{M' \times 3}, \hat{\mathbf{X}} \in \mathbb{R}^{N' \times 3}$ of the keypoints in the other point cloud (Sec. 4.3). The correspondences can then be obtained from the rows of $\tilde{\mathbf{X}}$ and $\hat{\mathbf{Y}}$, i.e. $\{\tilde{x}_i \leftrightarrow \hat{y}_i\}$ and similarly for the other direction. Concurrently, our network outputs overlap scores $\hat{o}_X \in \mathbb{R}^{M' \times 1}, \hat{o}_Y \in \mathbb{R}^{N' \times 1}$ that indicate the probability of each keypoint lying in the overlapping region. Finally, the required rigid transformation can be estimated directly from the correspondences within the overlap region (Sec. 4.4).

4.1. Downsampling and Feature Extraction

We follow [2, 23] to adopt the Kernel Point Convolution (KPConv) [45] backbone for feature extraction. The

KPConv backbone uses a series of ResNet-like blocks and strided convolutions to transform each input point cloud into a reduced set of keypoints $\tilde{\mathbf{X}} \in \mathbb{R}^{M' \times 3}, \tilde{\mathbf{Y}} \in \mathbb{R}^{N' \times 3}$ and their associated features $\mathbf{F}_{\tilde{\mathbf{X}}} \in \mathbb{R}^{M' \times D}, \mathbf{F}_{\tilde{\mathbf{Y}}} \in \mathbb{R}^{N' \times D}$. In contrast to [2, 23] that subsequently perform upsampling to obtain feature descriptors of the original point cloud resolution, our approach directly predicts the transformed keypoint locations using the downsampled features.

4.2. Transformer Cross-Encoder

The KPConv features from the previous step are linearly projected into a lower dimension $d = 256$. These projected features are then fed into $L = 6$ transformer cross-encoder¹ layers. Each transformer cross-encoder layer has three sub-layers: 1) a multi-head self-attention layer operating on the two point clouds separately, 2) a multi-head cross-attention layer which updates the features using information from the other point cloud, and 3) a position-wise feed-forward network. The cross-attention enables the network to compare points from the two different point clouds, and the self-attention allows points to interact with other points within the same point cloud when predicting its own transformed position, e.g. using rigidity constraints. Note that the network weights are shared among the two point clouds but not among the layers.

Attention sub-layers. The multi-head attention [49] operation in each sub-layer is defined as follows:

$$\text{MHAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = (\text{Head}_1 \oplus \dots \oplus \text{Head}_H) \mathbf{W}^O \quad (1a)$$

¹We name the layers cross-encoder layers to differentiate them from the usual transformer encoder layers [49] which only take in a single source.

$$\text{Head}_h = \text{Attn} \left(\mathbf{QW}_h^Q, \mathbf{KW}_h^K, \mathbf{VW}_h^V \right), \quad (1b)$$

where \oplus denotes concatenation over the channel dimension, $\mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V \in \mathbb{R}^{d \times d_{\text{head}}}$ and $\mathbf{W}^O \in \mathbb{R}^{Hd_{\text{head}} \times d}$ are learned projection matrices. We set the number of heads H to 8, and $d_{\text{head}} = d/H$. Each attention head employs a single-head dot product attention:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_{\text{head}}}} \right) \mathbf{V}. \quad (2)$$

Residual connections and layer normalization are applied to each sub-layer, and we use the “pre-LN” [54] ordering which we find to be easier to optimize.

The query, key, values are set to the same point cloud in the self-attention layers, *i.e.* $\text{MHAttn}(\mathbf{F}_{\tilde{X}}, \mathbf{F}_{\tilde{X}}, \mathbf{F}_{\tilde{X}})$ for the source point cloud (and likewise for the target point cloud). This allows points to attend to other parts within the same point cloud. For the cross-attention layers, the keys and values are set to be the features from the other point cloud, *i.e.* $\text{MHAttn}(\mathbf{F}_{\tilde{X}}, \mathbf{F}_{\tilde{Y}}, \mathbf{F}_{\tilde{Y}})$ for the source point cloud (and likewise for the target point cloud) to allow each point to interact with points in the other point cloud.

Position-wise Feed-forward Network. This sub-layer operates on the features of each keypoint individually. Following its usual implementation [49], we use a two-layer feed-forward network with a ReLU activation function after the first layer. Similar to the attention sub-layers, residual connections and layer normalization are applied.

Positional encodings. Unlike previous works [23, 50] that use attentions to learn discriminative features, our transformer layers replace the role of RANSAC and thus requires information of the point positions. Specifically, we incorporate positional information by adding sinusoidal positional encodings [49] to the inputs at each transformer layer.

The outputs of the transformer cross-encoder layers are features $\bar{\mathbf{F}}_{\tilde{X}} \in \mathbb{R}^{M' \times d}, \bar{\mathbf{F}}_{\tilde{Y}} \in \mathbb{R}^{N' \times d}$ which are conditioned on the other point cloud.

4.3. Output Decoding

The conditioned features can now be used to predict the coordinates of transformed keypoints. To this end, we use a two-layer MLP to regress the required coordinates. Particularly, the corresponding locations $\hat{\mathbf{Y}} \in \mathbb{R}^{M' \times 3}$ of the source keypoints $\tilde{\mathbf{X}}$ in the target point cloud are given as:

$$\hat{\mathbf{Y}} = \text{ReLU}(\bar{\mathbf{F}}_{\tilde{X}} \mathbf{W}_1 + \mathbf{b}_1) \mathbf{W}_2 + \mathbf{b}_2, \quad (3)$$

where $\mathbf{W}_1, \mathbf{W}_2$ and $\mathbf{b}_1, \mathbf{b}_2$ are learnable weights and biases, respectively. We use the hat accent ($\hat{\cdot}$) to indicate predicted quantities. A similar procedure is used to obtain the predicted transformed locations $\hat{\mathbf{X}}$ of the target keypoints.

Alternatively, we also explore the use of a single-head attention layer (*cf.* Tab. 4), where the predicted locations $\hat{\mathbf{Y}}$ are a weighted sum of the target keypoint coordinates $\tilde{\mathbf{Y}}$:

$$\hat{\mathbf{Y}} = \text{Attn}(\bar{\mathbf{F}}_X \mathbf{W}_{\text{out}}^Q, \bar{\mathbf{F}}_Y \mathbf{W}_{\text{out}}^K, \tilde{\mathbf{Y}}), \quad (4)$$

where $\text{Attn}(\cdot)$ is defined previously in Eq. (2), and $\mathbf{W}_{\text{out}}^Q, \mathbf{W}_{\text{out}}^K \in \mathbb{R}^{d \times d}$ are learned projection matrices.

In parallel, we separately predict the overlap confidence $\hat{\mathbf{o}}_X \in \mathbb{R}^{M' \times 1}, \hat{\mathbf{o}}_Y \in \mathbb{R}^{N' \times 1}$ using a single fully connected layer with sigmoid activation. This is used to mask out the influence of correspondences outside the overlap region which are not predicted as accurately.

4.4. Estimation of Rigid Transformation

The predicted transformed locations in both directions are first concatenated to obtain the final set of $M' + N'$ correspondences:

$$\hat{\mathbf{X}}_{\text{corr}} = \begin{bmatrix} \tilde{\mathbf{X}} \\ \hat{\mathbf{X}} \end{bmatrix}, \quad \hat{\mathbf{Y}}_{\text{corr}} = \begin{bmatrix} \tilde{\mathbf{Y}} \\ \hat{\mathbf{Y}} \end{bmatrix}, \quad \hat{\mathbf{o}}_{\text{corr}} = \begin{bmatrix} \hat{\mathbf{o}}_X \\ \hat{\mathbf{o}}_Y \end{bmatrix}. \quad (5)$$

The required rigid transformation can be estimated from the estimated correspondences by solving the following:

$$\hat{\mathbf{R}}, \hat{\mathbf{t}} = \arg \min_{\mathbf{R}, \mathbf{t}} \sum_i^{M'+N'} \hat{o}_i \|\mathbf{R}\hat{\mathbf{x}}_i + \mathbf{t} - \hat{\mathbf{y}}_i\|^2, \quad (6)$$

where $\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i, \hat{o}_i$ denote the i^{th} row of $\hat{\mathbf{X}}_{\text{corr}}, \hat{\mathbf{Y}}_{\text{corr}}, \hat{\mathbf{o}}_{\text{corr}}$, respectively. We follow [20, 57] to solve Eq. (6) in closed form using a weighted variant of the Kabsch-Umeyama [25, 47] algorithm.

4.5. Loss Functions

We train our network end-to-end with the ground truth poses $\{\mathbf{R}^*, \mathbf{t}^*\}$ for supervision using the following losses:

Overlap loss. The predicted overlap scores are supervised using the binary cross entropy loss. The loss for the source point cloud \mathbf{X} is given by:

$$\mathcal{L}_o^X = -\frac{1}{M'} \sum_i^{M'} o_{\tilde{\mathbf{x}}_i}^* \cdot \log \hat{o}_{\tilde{\mathbf{x}}_i} + (1 - o_{\tilde{\mathbf{x}}_i}^*) \cdot \log (1 - \hat{o}_{\tilde{\mathbf{x}}_i}). \quad (7)$$

To obtain the ground truth overlap labels $o_{\tilde{\mathbf{x}}_i}^*$, we first compute the dense ground truth labels for the original point cloud in a similar fashion as [23]. Specifically, the ground truth label for point $\mathbf{x}_i \in \mathbf{X}$ is defined as:

$$o_{\mathbf{x}_i}^* = \begin{cases} 1, & \|\mathcal{T}^*(\mathbf{x}_i) - \text{NN}(\mathcal{T}^*(\mathbf{x}_i), \mathbf{Y})\| < r_o \\ 0, & \text{otherwise} \end{cases}, \quad (8)$$

where $\mathcal{T}^*(\mathbf{x}_i)$ denotes the application of the ground truth rigid transform $\{\mathbf{R}^*, \mathbf{t}^*\}$, $\text{NN}(\cdot)$ denotes the spatial nearest neighbor and r_o is a predefined overlap threshold. We

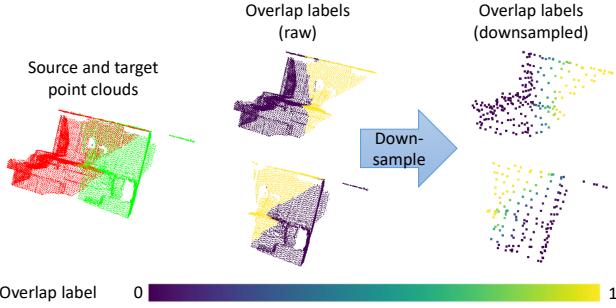


Figure 3. Pair of point clouds (left) and their corresponding ground truth overlap labels for the dense points (middle) and downsampled keypoints (right). Note that the keypoints near the overlap boundaries have a ground truth label between 0 and 1.

then obtain the overlap labels $o_{\tilde{\mathbf{x}}_i}^*$ for the downsampled keypoints through average pooling using the same pooling indices from the KPConv downsampling. See Fig. 3 for an example of our overlap ground truth labels. The loss \mathcal{L}_o^Y for the target point cloud \mathbf{Y} is obtained in a similar fashion. We thus get a total overlap loss of: $\mathcal{L}_o = \mathcal{L}_o^X + \mathcal{L}_o^Y$.

Correspondence loss. We apply a ℓ^1 loss on the predicted transformed locations for keypoints in the overlapping region:

$$\mathcal{L}_c^X = \frac{1}{\sum_i o_{\tilde{\mathbf{x}}_i}^*} \sum_i o_{\tilde{\mathbf{x}}_i}^* |\mathcal{T}^*(\tilde{\mathbf{x}}_i) - \hat{\mathbf{y}}_i|, \quad (9)$$

and similarly for the target point cloud. We thus get a total correspondence loss of: $\mathcal{L}_c = \mathcal{L}_c^X + \mathcal{L}_c^Y$.

Feature loss. To encourage the network to take into account geometric properties when computing the correspondences, we apply an InfoNCE [35] loss on the conditioned features. Considering the set of points $\mathbf{x} \in \tilde{\mathbf{X}}$ with a correspondence in $\tilde{\mathbf{Y}}$, the InfoNCE loss for the source point cloud is:

$$\mathcal{L}_f^X = -\mathbb{E}_{\mathbf{x} \in \tilde{\mathbf{X}}} \left[\log \frac{f(\mathbf{x}, \mathbf{p}_x)}{f(\mathbf{x}, \mathbf{p}_x) + \sum_{\mathbf{n}_x} f(\mathbf{x}, \mathbf{n}_x)} \right], \quad (10)$$

where we follow [35] to use a log-bilinear model for $f(\cdot, \cdot)$:

$$f(\mathbf{x}, \mathbf{c}) = \exp(\bar{\mathbf{f}}_x^T \mathbf{W}_f \bar{\mathbf{f}}_c). \quad (11)$$

$\bar{\mathbf{f}}_x$ denotes the conditioned feature for point \mathbf{x} . \mathbf{p}_x and \mathbf{n}_x denote keypoints in $\tilde{\mathbf{Y}}$ which match and do not match \mathbf{x} , respectively. They are determined using the positive and negative margins (r_p, r_n) which are set as $(m, 2m)$. m is the voxel distance used in the final downsampling layer in the KPConv backbone, and all negative points that falls outside the negative margin are utilized for \mathbf{n}_x . Since the two point clouds contain the same type of features, we enforce the learnable linear transformation \mathbf{W}_f to be symmetrical by parameterizing it as the sum of a upper triangular matrix

\mathbf{U}_f and its transpose, i.e. $\mathbf{W}_f = \mathbf{U}_f + \mathbf{U}_f^\top$. As explained in [35], Eq. (10) maximizes the mutual information between features for matching points. Unlike [2, 23], we do not use the circle loss [44] that requires the matching features to be similar (w.r.t. ℓ^2 or cosine distance). This is unsuitable since: 1) our conditioned features contains information about the transformed positions, and 2) our keypoints are sparse and are unlikely to be at the same location in the two point clouds, and therefore the geometric features are also different.

Our final loss is a weighted sum of the three components: $\mathcal{L} = \mathcal{L}_c + \lambda_o \mathcal{L}_o + \lambda_f \mathcal{L}_f$, where we set $\lambda_o = 1.0$ and $\lambda_f = 0.1$ for all experiments.

5. Experiments

5.1. Implementation Details

We train our network using AdamW [33] optimizer with a initial learning rate of 0.0001 and weight decay of 0.0001. Gradients are clipped at 0.1. For the 3DMatch dataset, we train for 60 epochs with a batch size of 2, halving the learning rate every 20 epochs. We train on the ModelNet40 dataset for 400 epochs with a batch size of 4, and halving the learning rate every 100 epochs. Training requires around 2.5 and 2 days for 3DMatch and ModelNet40 on a single Nvidia Titan RTX, respectively.

5.2. Datasets and Results

3DMatch. The 3DMatch dataset [61] contains 46 train, 8 validation and 8 test scenes. We use the preprocessed data from [23] containing voxel-grid downsampled point clouds, and follow them to evaluate on both pairs with $> 30\%$ overlap (3DMatch) and $10 - 30\%$ overlap (3DLoMatch). Each input point cloud contains an average of about 20,000 points, which are downsampled to an average of 345 points by our KPConv backbone. We perform training data augmentation by applying small rigid perturbations, jittering of the point locations and shuffling of points.

Following [2, 10, 23], we evaluate using Registration Recall (RR) which measures the fraction of successfully registered pairs, defined as having a correspondence RMSE below 0.2m. We also evaluate on the Relative Rotation Errors (RRE) and Relative Translation Errors (RTE) that measures the accuracy of successful registrations. We follow [23] and compare against several recent learned correspondence-based algorithms [2, 14, 21, 23]². These algorithms tend to perform better with a larger number of sampled interest points, and therefore we only show the results for the maximum number (5000) of sampled points. Since Predator [23] obtains the highest registration recall when using 1000 interest points, we also reran their open-source

²The initial Predator code had a bug which decreased the performance, and we list the improved results using its corrected version.

Method	3DMatch			3DLoMatch		
	RR(%)	RRE($^{\circ}$)	RTE(m)	RR(%)	RRE($^{\circ}$)	RTE(m)
3DSN [21]	78.4	2.199	0.071	33.0	3.528	0.103
FCGF [14]	85.1	1.949	0.066	40.1	3.147	0.100
D3Feat [2]	81.6	2.161	0.067	37.2	3.361	0.103
Predator-5k [23]	89.0	2.029	0.064	59.8	3.048	0.093
Predator-1k [23]	90.5	2.062	0.068	62.5	3.159	0.096
Predator-NR [23]	62.7	2.582	0.075	24.0	5.886	0.148
OMNet [55]	35.9	4.166	0.105	8.4	7.299	0.151
DGR [13]	85.3	2.103	0.067	48.7	3.954	0.113
PCAM [6]	85.5	1.808	0.059	54.9	3.529	0.099
Ours	92.0	1.567	0.049	64.8	2.827	0.077

Table 1. Performance on 3DMatch and 3DLoMatch datasets. Results for 3DSN, FCGF, D3Feat and Predator-5k are from [23].

Methods	ModelNet			ModelLoNet		
	RRE	RTE	CD	RRE	RTE	CD
PointNetLK [1]	29.725	0.297	0.0235	48.567	0.507	0.0367
OMNet [55]	2.947	0.032	0.0015	6.517	0.129	0.0074
DCP-v2 [50]	11.975	0.171	0.0117	16.501	0.300	0.0268
RPM-Net [57]	<u>1.712</u>	<u>0.018</u>	<u>0.00085</u>	7.342	<u>0.124</u>	<u>0.0050</u>
Predator [23]	1.739	0.019	0.00089	<u>5.235</u>	0.132	0.0083
Ours	1.473	0.014	0.00078	3.930	0.087	0.0037

Table 2. Evaluation results on ModelNet40 dataset. Results of DCP-v2, RPM-Net and Predator are taken are from [23].

code with 1000 interest points and include the results under Predator-1k. Furthermore, we compare with several methods [6, 13, 55] designed to avoid RANSAC. We trained OMNet [55] on 3DMatch with a batch size of 32 for 2000 epochs using 1024 random points. For [6, 13], we use the authors’ trained weights. We disabled ICP refinement in DGR [13] for a fair comparison.

Table 1 shows the quantitative results, and Figs. 1 and 8a to 8d show several examples of the qualitative results. We also show the results for the individual scenes in the supplementary. For both 3DMatch and 3DLoMatch benchmarks, our method achieves the highest average registration recall across scenes. Interestingly, our registration is also very precise and achieved the lowest RTE and RRE on both 3DMatch and 3DLoMatch benchmarks despite only using a small number of points for pose estimation. In addition, we also compare with Predator-NR, a RANSAC-free variant of Predator-1k that utilizes the product of the predicted overlap and matchability scores to weigh the correspondences during pose estimation. The underperformance of Predator-NR indicates that our proposed method is more suitable for replacing RANSAC. The results support our claim that our attention mechanism can replace the role of RANSAC since our REGTR uses largely the same KPConv backbone as Predator. Lastly, we note that OMNet does not perform well on the 3DMatch dataset. This is likely due to the difficulty in describing complex scenes with a single global feature vector. This behavior is also previously observed in [13] for another global feature-based algorithm, PointNetLK [1].

Method	Preproc.	Feat. Extract.	Pose est.	Total
3DSN [21]	27938	1872	2588	32398
FCGF [14]	15	41	1597	1653
D3Feat [2]	174*	27	795	996
Predator-5k [23]	245*	45	1017	1306
Predator-1k [23]	245*	45	189	479
OMNet [55]	—	9	1	10
DGR [13]	28	31	1258	1318
PCAM [6]	—	520	1063	1584
Ours	35*	54	2	91

Table 3. Run time in milliseconds on the 3DMatch benchmark test set. *Similar pre-processing is used for D3Feat, Predator and our algorithm, but our algorithm uses a faster GPU implementation.

ModelNet40. We also evaluate on the ModelNet40 [52] dataset comprising synthetic CAD models. We follow the data setting in [23, 57], where the point clouds are sampled randomly from mesh faces of the CAD models, cropped and subsampled. Following [23], we evaluate on two partial overlap settings: *ModelNet* which has 73.5% pairwise overlap on average, and *ModelLoNet* which contains a lower 53.6% average overlap. We train only on ModelNet, and perform direct generalization to ModelLoNet. We follow [23, 57] and measure the performance using Relative Rotation Error (RRE) and Relative Translation Error (RTE) on all point clouds, as well as the Chamfer distance (CD) between the registered scans.

The results are shown in Tab. 2, with example qualitative results in Figs. 8e and 8f. We compare against recent correspondence-based [23] and end-to-end registration methods [1, 50, 55, 57]. Predator [23] samples 450 points in this experiment. OMNet [55] was originally trained only on axis-asymmetrical categories, and we retrained it on all categories to obtain a slightly improved result. As noted in [23], many of the end-to-end registration methods are specifically tuned for ModelNet. RPM-Net [57] additionally uses surface normal information. Despite this, our REGTR substantially outperforms all baseline methods in all metrics under both normal overlap (ModelNet) and low overlap (ModelLoNet) regimes. Our learned attention mechanism is able to outperform the optimal transport (in RPM-Net) and RANSAC step (in Predator).

5.3. Analysis

We perform further analysis in this section to better understand our algorithm behavior. All experiments in this section are performed on the larger 3DMatch dataset.

Runtime. We compare the runtime of REGTR against several algorithms in Tab. 3. We conducted the test on a single Nvidia Titan RTX with Intel Core i7-6950X @ 3.0GHz and 64GB RAM. Our entire pipeline runs under 100ms, and is feasible for many real-time applications. The time consuming step for correspondence-based algorithms

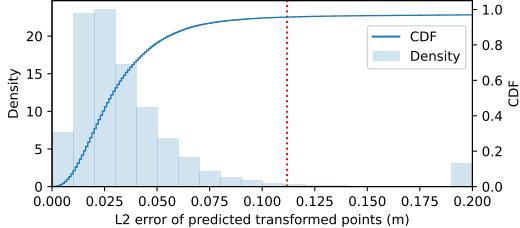


Figure 4. Histogram and CDF plot of ℓ^2 errors of predicted correspondences. 95.7% of predicted correspondences have errors below 0.112m, the median keypoint-to-keypoint distance (denoted by red dashed line). Errors are clipped at 0.2m for clarity, with only 3.0% of predicted correspondences exceeding this error.

is the pose estimation which includes feature matching and RANSAC. For example, excluding preprocessing required for the KPConv backbone, Predator [23] takes 234ms for the registration when sampling just 1,000 points. DGR [13] and PCAM [6] also require long times for pose estimation due to their robust refinement and RANSAC safeguard. Although the global feature-based OMNet runs faster than our algorithm, we note that it is unable to obtain good accuracy on the 3DMatch dataset.

Accuracy of predicted correspondences. In Fig. 4, we plot the distribution of ℓ^2 -error of the predicted correspondences for keypoints within the overlap region (where $o_{\tilde{\mathbf{x}}_i}^*, o_{\tilde{\mathbf{y}}_i}^* > 0.5$) for the 3DMatch test set. The median error of our predicted correspondences is 0.028m, which is significantly smaller than the median distance between keypoints (0.112m). For comparison, an oracle matcher that matches every keypoint to the closest keypoint using the ground truth pose obtains a median error of 0.071m. Our direct prediction of correspondences is able to overcome the resolution issues from the downsampling, and thus explains the precise registration obtained by REGTR.

We also visualize the predicted correspondences of a point cloud pair in Fig. 5. The short green error lines in Fig. 5b indicate our predicted transformed locations are highly accurate within the overlap region, even in non-informative regions (*e.g.* floor). Interestingly, correspondence for points outside the overlap region are projected to near the overlap boundaries. These observations suggest that REGTR is able to make use of rigidity constraints to guide the positions of the predicted correspondences.

Visualization of attention. In Fig. 6, we visualize the attention for a point on the ground for the same point cloud pair from the previous section. Since the point lies in a non-informative region, the point attends to multiple similar looking regions in the other point cloud in the first transformer layer (Fig. 6a). At the sixth layer, the point is confident of its position and mostly focuses on its correct corresponding location (Fig. 6b). The self-attention in Fig. 6c

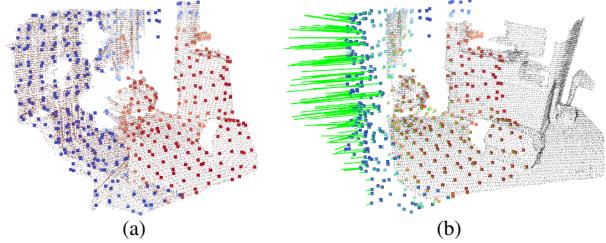


Figure 5. Visualization of predicted correspondences. Keypoints are colored based on their predicted overlap score, where high scores are denoted in red. (a) Source \mathbf{X} and keypoints $\tilde{\mathbf{X}}$, (b) Target \mathbf{Y} and predicted correspondences $\tilde{\mathbf{Y}}$, with green lines showing the correspondence error. Best viewed in color.

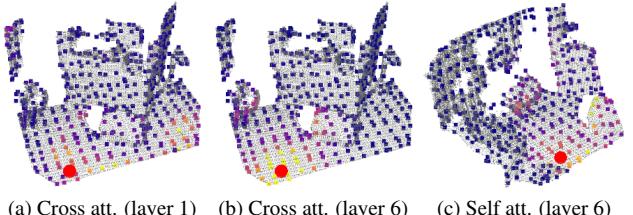


Figure 6. Visualization of attention weights for the point indicated with a red dot. Brighter colors indicate higher attention.

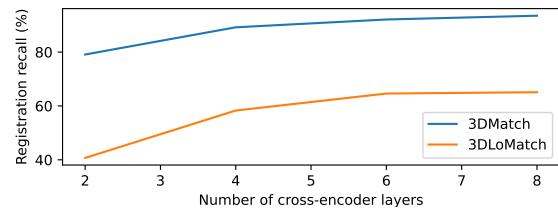


Figure 7. Performance for various number of cross-encoder layers.

shows that the point makes use of feature rich regions within the same point cloud to help localize its correct location.

5.4. Ablations

We further perform ablation studies on the 3DMatch dataset to understand the role of various components.

Number of cross-encoder layers. We evaluate how the performance varies with the number of cross-encoder layers in Fig. 7. Our network cannot function without any cross-encoder layers, and we show the registration recall for two to eight layers. Performance generally improves with more cross-encoder layers, but saturates around $L = 6$ encoder layers (which we use for all our experiments).

Comparison with RANSAC. We compare with a version of REGTR where we replace our output decoder with a two-layer MLP which outputs 256D feature descriptors and a parallel single-layer decoder that outputs the overlap score. The pose is subsequently estimated using RANSAC on nearest neighbor feature matches. The network is trained

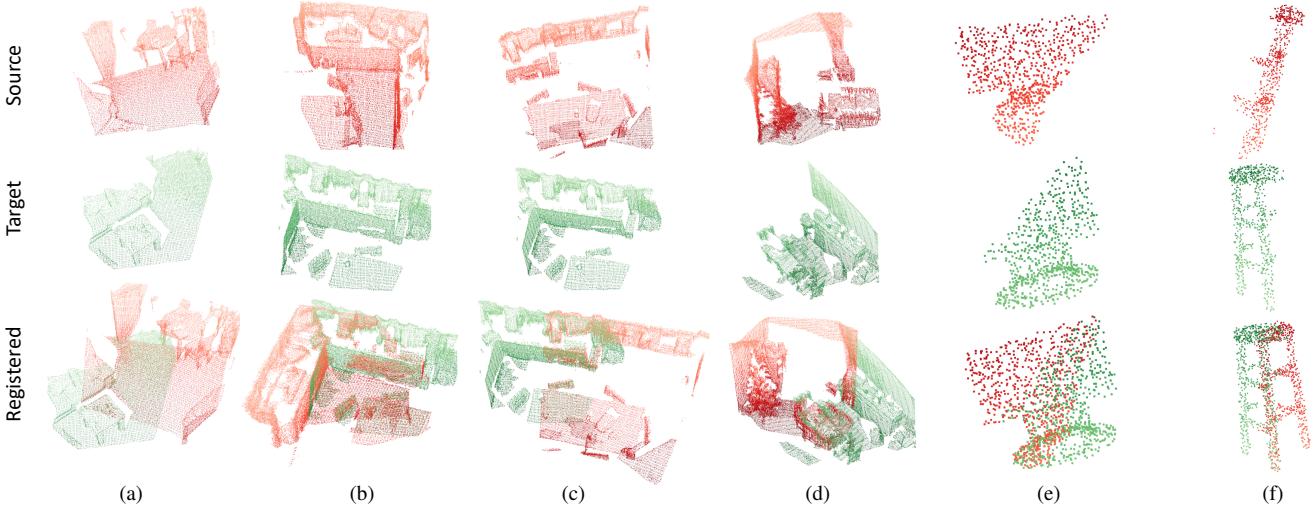


Figure 8. Example qualitative registration results for (a, b) 3DMatch, (c, d) 3DLoMatch (e) ModelNet40, and (f) ModelLoNet.

Method	3DMatch			3DLoMatch		
	RR(%)	RRE($^{\circ}$)	RTE(m)	RR(%)	RTE($^{\circ}$)	RTE(m)
RANSAC baseline	87.7	2.296	0.072	52.7	4.038	0.112
Weighted coor.	90.9	1.468	0.046	63.1	2.540	0.076
No feature loss	90.4	1.638	0.049	61.9	2.898	0.083
Circle loss	90.0	1.696	0.051	61.2	3.217	0.092
Loss on all layers	83.9	1.781	0.053	46.8	3.448	0.101
REGTR	92.0	<u>1.567</u>	<u>0.049</u>	64.8	2.827	<u>0.077</u>
REGTR+RANSAC	<u>91.9</u>	1.607	<u>0.049</u>	63.3	<u>2.753</u>	0.079

Table 4. Ablation of components and losses

using only \mathcal{L}_o and \mathcal{L}_f (using Circle Loss [44]). This results in a lower registration recall, and significantly higher rotation and translation errors as the downsampled keypoints do not provide enough resolution for accurate registration.

We also try applying RANSAC to the predicted correspondences from REGTR to see if the performance can further improve. Row 7 of Tab. 4 shows marginally worse registration recall. This indicates that RANSAC is no longer beneficial on the predicted correspondences that are already consistent with a rigid transformation.

Decoding scheme. We compare with decoding the coordinates as a weighted sum of coordinates (Eq. 4). Compared to our simpler approach of regressing the coordinates using a MLP, computing the coordinates as a weighted sum achieves a slightly better RTE and RRE, but lower registration recall. See rows 2 and 6 of Tab. 4.

Loss ablations. Rows 3-6 of Tab. 4 shows the registration performance with different loss configurations. Without the feature loss to guide the network outputs, the network obtained a 1.6% and 2.9% lower registration recall for 3DMatch and 3DLoMatch, respectively. Using circle loss from [23] also underperformed as the network cannot incorporate positional information into the feature as effectively. We also experimented with applying the losses on all $L = 6$

transformer layers (instead of just the final one) with the output decoder shared among all cross-encoder layers. This additional supervision led to a 8.1% (3DMatch) and 18.0% (3DLoMatch) lower registration recall. Consequently, we only apply the supervision for the output of the last cross-encoder layer.

6. Limitations

Our use of transformers layers with quadratic complexity prevents its use on large number of points, and we can only apply them on downsampled point clouds. Although our direct correspondence prediction alleviates the resolution issue, it is possible that a finer resolution can result in even higher performance. We have tried transformer layers with linear complexity [12, 26], but that obtained subpar performance. Alternate workarounds include using sparse attention [9], or performing a coarse-to-fine registration.

7. Conclusions

We propose the REGTR for rigid point cloud registration, which directly predicts clean point correspondences using multiple transformer layers. The rigid transformation can then be estimated from the correspondences without further nearest neighbor feature matching nor RANSAC steps. The direct prediction of correspondences overcomes the resolution issues from the use of downsampled features, and our method achieves state-of-the-art performance on both scene and object point cloud datasets.

Acknowledgement. This research/project is supported in part by the National Research Foundation, Singapore under its AI Singapore Program (AISG Award No: AISG2-RP-2020-016), and the Tier 2 grant MOE-T2EP20120-0011 from the Singapore Ministry of Education.

References

- [1] Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivastan, and Simon Lucey. Pointnetlk: Robust & efficient point cloud registration using pointnet. In *CVPR*, pages 7163–7172, 2019. [2](#), [6](#)
- [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*, pages 6359–6367, 2020. [1](#), [2](#), [3](#), [5](#), [6](#), [15](#)
- [3] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *IJCV*, 56(3):221–255, 2004. [2](#)
- [4] Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE TPAMI*, 14(2):239–256, 1992. [2](#)
- [5] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. DSAC - Differentiable RANSAC for camera localization. In *CVPR*, pages 6684–6692, 2017. [2](#)
- [6] Anh-Quan Cao, Gilles Puy, Alexandre Boulch, and Renaud Marlet. PCAM: Product of cross-attention matrices for rigid registration of point clouds. In *ICCV*, pages 13229–13238, 2021. [2](#), [6](#), [7](#), [15](#)
- [7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. [1](#), [3](#)
- [8] Y. Chen and G. Medioni. Object modeling by registration of multiple range images. In *ICRA*, pages 2724–2729 vol.3, 1991. [2](#)
- [9] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. [8](#)
- [10] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *CVPR*, pages 5556–5565, 2015. [5](#)
- [11] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, pages 539–546, 2005. [2](#)
- [12] Krzysztof Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. In *ICLR*, 2021. [8](#)
- [13] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *CVPR*, pages 2514–2523, 2020. [2](#), [6](#), [7](#), [15](#)
- [14] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *ICCV*, pages 8958–8966, 2019. [1](#), [2](#), [5](#), [6](#), [15](#)
- [15] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH*, pages 303–312, 1996. [11](#)
- [16] Angela Dai, Matthias Nießner, Michael Zollöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface re-integration. *ACM TOG*, 2017. [11](#)
- [17] Haowen Deng, Tolga Birdal, and Slobodan Ilic. PPFNet: Global context aware local features for robust 3d point matching. In *CVPR*, pages 195–205, 2018. [1](#), [2](#)
- [18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. [3](#)
- [19] Kai Fischer, Martin Simon, Florian Olsner, Stefan Milz, Horst-Michael Groß, and Patrick Mader. Stickypillars: Robust and efficient feature matching on point clouds using graph neural networks. In *CVPR*, pages 313–323, 2021. [2](#)
- [20] Zan Gojcic, Caifa Zhou, Jan D Wegner, Leonidas J Guibas, and Tolga Birdal. Learning multiview 3d point cloud registration. In *CVPR*, pages 1759–1769, 2020. [2](#), [4](#)
- [21] 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. [5](#), [6](#), [15](#)
- [22] Maciej Halber and Thomas Funkhouser. Fine-to-coarse global registration of rgb-d scans. In *CVPR*, pages 1755–1764, 2017. [11](#)
- [23] Shengyu Huang, Zan Gojcic, Mikhail Usyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *CVPR*, pages 4267–4276, 2021. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [11](#), [15](#)
- [24] 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. [2](#)
- [25] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976. [4](#), [11](#)
- [26] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, pages 5156–5165, 2020. [8](#)
- [27] Marc Khoury, Qian-Yi Zhou, and Vladlen Koltun. Learning compact geometric features. In *Proceedings of the IEEE international conference on computer vision*, pages 153–161, 2017. [2](#)
- [28] Kevin Lai, Liefeng Bo, and Dieter Fox. Unsupervised feature learning for 3d scene labeling. In *ICRA*, pages 3050–3057, 2014. [11](#)
- [29] Junha Lee, Seungwook Kim, Minsu Cho, and Jaesik Park. Deep hough voting for robust global registration. In *ICCV*, pages 15994–16003, 2021. [2](#)
- [30] Jiaxin Li and Gim Hee Lee. USIP: Unsupervised stable interest point detection from 3d point clouds. In *ICCV*, pages 361–370, 2019. [2](#)
- [31] Jiahao Li, Changhao Zhang, Ziyao Xu, Hangning Zhou, and Chi Zhang. Iterative distance-aware similarity matrix convolution with mutual-supervised point elimination for efficient point cloud registration. In *ECCV*, pages 378–394, 2020. [2](#)
- [32] Xueqian Li, Jhony Kaesemodel Pontes, and Simon Lucey. Pointnetlk revisited. In *CVPR*, pages 12763–12772, 2021. [2](#)
- [33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. [5](#)

- [34] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *ICCV*, pages 2906–2917, 2021. [1](#), [3](#)
- [35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. [5](#)
- [36] G Dias Pais, Srikumar Ramalingam, Venu Madhav Govindu, Jacinto C Nascimento, Rama Chellappa, and Pedro Miraldo. 3DRegNet: A deep neural network for 3D point registration. In *CVPR*, pages 7193–7203, 2020. [2](#)
- [37] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. [2](#), [11](#)
- [38] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In *ICRA*, pages 3212–3217, 2009. [2](#)
- [39] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. [1](#)
- [40] Vinit Sarode, Xueqian Li, Hunter Goforth, Yasuhiro Aoki, Rangaprasad Arun Srivatsan, Simon Lucey, and Howie Choset. Pcrnet: Point cloud registration network using pointnet encoding. *arXiv preprint arXiv:1908.07906*, 2019. [2](#)
- [41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. [2](#)
- [42] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgbd images. In *CVPR*, pages 2930–2937, 2013. [11](#)
- [43] Bastian Steder, Radu Bogdan Rusu, Kurt Konolige, and Wolfram Burgard. NARF: 3D range image features for object recognition. In *Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics at the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, volume 44, 2010. [2](#)
- [44] Yifan Sun, Changmao Cheng, Yuhang Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, pages 6398–6407, 2020. [5](#), [8](#)
- [45] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *ICCV*, pages 6411–6420, 2019. [2](#), [3](#), [12](#)
- [46] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique shape context for 3d data description. In *ACM Workshop on 3D Object Retrieval*, pages 57–62, 2010. [2](#)
- [47] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE TPAMI*, 13(04):376–380, 1991. [4](#)
- [48] Julien Valentin, Angela Dai, Matthias Niessner, Pushmeet Kohli, Philip Torr, Shahram Izadi, and Cem Keskin. Learning to navigate the energy landscape. In *3DV*, pages 323–332, 2016. [11](#)
- [49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. [1](#), [2](#), [3](#), [4](#), [11](#)
- [50] Yue Wang and Justin M. Solomon. Deep closest point: Learning representations for point cloud registration. In *ICCV*, pages 3523–3532, 2019. [1](#), [2](#), [4](#), [6](#), [11](#)
- [51] Yue Wang and Justin M Solomon. Prnet: Self-supervised learning for partial-to-partial registration. *NeurIPS*, 32, 2019. [1](#), [2](#)
- [52] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Lin-guang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920, 2015. [2](#), [6](#), [11](#)
- [53] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. SUN3D: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, pages 1625–1632, 2013. [11](#)
- [54] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *ICML*, pages 10524–10533, 2020. [4](#), [12](#)
- [55] Hao Xu, Shuaicheng Liu, Guangfu Wang, Guanghui Liu, and Bing Zeng. OMNet: Learning overlapping mask for partial-to-partial point cloud registration. In *ICCV*, pages 3132–3141, 2021. [2](#), [6](#), [15](#)
- [56] Zi Jian Yew and Gim Hee Lee. 3DFeat-Net: Weakly supervised local 3d features for point cloud registration. In *ECCV*, pages 607–623, 2018. [1](#), [2](#)
- [57] Zi Jian Yew and Gim Hee Lee. RPM-Net: Robust point matching using learned features. In *CVPR*, pages 11824–11833, 2020. [1](#), [2](#), [4](#), [6](#), [11](#)
- [58] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *CVPR*, pages 2666–2674, 2018. [2](#)
- [59] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *ICCV*, pages 12498–12507, 2021. [3](#)
- [60] Wentao Yuan, Benjamin Eckart, Kihwan Kim, Varun Jampani, Dieter Fox, and Jan Kautz. DeepGMR: Learning latent gaussian mixture models for registration. In *ECCV*, pages 733–750, 2020. [2](#)
- [61] 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. [1](#), [2](#), [5](#), [11](#)
- [62] Yu Zhong. Intrinsic shape signatures: A shape descriptor for 3d object recognition. In *ICCV Workshops*, pages 689–696, 2009. [2](#)

Supplementary Material

In this supplementary, we first provide additional details on the datasets and their preprocessing (Sec. A). We then describe the procedure for recovering the rigid transformation from the correspondences (Sec. B), our sinusoidal position encodings (Sec. C) and additional information on the network architecture (Sec. D). Finally, we show detailed results for ScanNet (Sec. E) and additional qualitative results (Sec. F).

A. Dataset Details

3DMatch. The 3DMatch [61] dataset comprises RGB-D frames obtained from several sources (Tab. 5). The data is captured from diverse scenes (*e.g.* bedrooms, kitchens, offices) and different sensors (*e.g.* Microsoft Kinect, Intel Realsense), and each point cloud is generated by fusing 50 consecutive depth frames using TSDF volumetric fusion [15]. We use the voxel-grid down-sampled data from Predator [23], and the same point cloud pairs for training and evaluation. The dataset contains 46 train, 8 validation, and 8 test scenes. The training and validation scenes contains a total of 20,586³ and 1,331 point cloud pairs respectively, and the test scenes contain 1,279 (3DMatch) and 1,726 (3DLoMatch) pairs. We apply training data augmentation by applying a small rigid perturbation with magnitudes sampled from a Gaussian distribution with $\sigma_r = 0.1\pi$ and $\sigma_t = 0.1$ for the rotation and translation, respectively. We then apply a Gaussian noise ($\sigma = 0.05$) on the individual point locations, and shuffling of point order.

ModelNet40. The ModelNet40 [52] dataset provides 3D CAD models from 40 object categories for academic use. We follow previous works [50, 57] in using the preprocessed data from [37]. These data are generated by sampling 2,048 points from the mesh faces and then scaling them to fit into a unit sphere. The partial scans are generated from the procedure in [57]: A half-space with random direction is sampled, and shifted such that a proportion p of points lie within the half space. Subsequently, random rotation of up to 45°, translation up to 0.5 units, Gaussian noise ($\sigma = 0.05$) on the individual point locations, and shuffling of point order are applied to the point clouds. The point clouds are finally resampled to 717 points. Following [23], p is set to 0.7 and 0.5 for ModelNet and ModelLoNet benchmarks, respectively. We use the first 20 categories for training and validation, and the other 20 categories for testing.

Datasets	License
SUN3D [22, 53]	CC BY-NC-SA 4.0
7-Scenes [42]	Non-commercial use only (License not stated)
RGB-D Scenes v2 [28]	CC BY-NC-SA 4.0
BundleFusion [16]	CC BY-NC-SA 4.0
Analysis-by-Synthesis [48]	CC BY-NC-SA 4.0

Table 5. Raw data used in the 3DMatch [61] dataset and their licenses.

³one point cloud (7-scenes-fire/19) has a wrong groundtruth pose and we exclude training pairs containing this point cloud.

B. Estimation of Rigid Transformation

In this section, we describe the closed form solution for the rigid transformation $\{\mathbf{R}, \mathbf{t}\}$, given correspondences $\{\mathbf{x}_i \leftrightarrow \mathbf{y}_i\}$ with their weights $\{o_i\}$, as used in Sec. 4.4:

$$\hat{\mathbf{R}}, \hat{\mathbf{t}} = \arg \min_{\mathbf{R}, \mathbf{t}} \sum_i^N o_i \|\mathbf{R}\mathbf{x}_i + \mathbf{t} - \mathbf{y}_i\|^2. \quad (12)$$

Step 1. Compute the weighted centroids of the 2 point sets:

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^N o_i \mathbf{x}_i}{\sum_{i=1}^N o_i}, \quad \bar{\mathbf{y}} = \frac{\sum_{i=1}^N o_i \mathbf{y}_i}{\sum_{i=1}^N o_i}. \quad (13)$$

Step 2. Center the point clouds by subtracting away the centroid:

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}, \quad \tilde{\mathbf{y}}_i = \mathbf{y}_i - \bar{\mathbf{y}}, \quad \forall i = 1, \dots, N. \quad (14)$$

Step 3. Recover the rotation \mathbf{R} . For this, we can use the Kabisch algorithm [25]. First construct the following 3×3 weighted covariance matrix:

$$\mathbf{H} = \sum_{i=1}^N o_i \tilde{\mathbf{x}}_i \tilde{\mathbf{y}}_i^\top. \quad (15)$$

Considering the singular value decomposition $\mathbf{H} = \mathbf{U}\Sigma\mathbf{V}^\top$, the desired rotation $\hat{\mathbf{R}}$ is given by:

$$\hat{\mathbf{R}} = \mathbf{V} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \det(\mathbf{V}\mathbf{U}^\top) \end{bmatrix} \mathbf{U}^\top, \quad (16)$$

where $\det(\cdot)$ denotes the matrix determinant.

Step 4. Lastly, the translation can be computed as:

$$\hat{\mathbf{t}} = \bar{\mathbf{y}} - \hat{\mathbf{R}}\bar{\mathbf{x}}. \quad (17)$$

C. Position Encodings

We encode the point coordinates by generalizing the sinusoidal positional encodings in [49] to 3D continuous coordinates. The position encodings have the same dimension $d = 256$ as the feature embeddings used in the attention layers.

For a point $\mathbf{x} = (x, y, z)$, we separately transform each coordinate to their embeddings $\mathbf{p}_x^\mathbf{x}, \mathbf{p}_y^\mathbf{x}, \mathbf{p}_z^\mathbf{x} \in \mathbb{R}^{2^{\lfloor d/6 \rfloor}}$. The x -coordinate is transformed as:

$$\mathbf{p}_x^\mathbf{x}[2i] = \sin\left(\frac{x}{10000^{2i/\lfloor d/3 \rfloor}}\right) \quad (18a)$$

$$\mathbf{p}_x^\mathbf{x}[2i+1] = \cos\left(\frac{x}{10000^{2i/\lfloor d/3 \rfloor}}\right). \quad (18b)$$

The y and z coordinates are transformed in a similar manner. We then concatenate the embeddings for all three dimensions. Since the embedding dimension $d = 256$ is not divisible by 6, we pad the remaining 4 elements with zeros to obtain the final embedding.

Dec.	Pos.	3DMatch			3DLoMatch		
		RR(%)	RRE($^{\circ}$)	RTE(m)	RR(%)	RRE($^{\circ}$)	RTE(m)
Wt.	None	87.9	1.958	0.062	55.0	3.389	0.093
Wt.	Learned	89.0	<u>1.519</u>	<u>0.047</u>	61.3	<u>2.812</u>	0.083
Wt.	Sine	90.9	1.468	0.046	<u>63.1</u>	2.540	0.076
Reg.	Learned	91.6	1.576	0.047	64.8	2.999	0.080
Reg.	Sine	92.0	1.567	0.049	64.8	<u>2.827</u>	<u>0.077</u>

Table 6. Effects of different position encodings. “Dec.” denotes correspondence decoding scheme, which can be either weighted coordinates using Eq. 4 (Wt.) or regression using Eq. 3 (Reg.). “Pos.” denotes position encoding type.

Importance of position encodings. Table 6 compares different choices of position encodings. The position encodings can only be removed when decoding the correspondences as a weighted sum (Eq. 4). Without position encodings, the network suffered a significant drop in performance both in terms of registration recall (RR) and accuracy (RTE and RRE), further supporting our hypothesis that our attention mechanism utilizes rigidity constraints to correct bad matches. We also compare with learned embeddings (using a 5-layer MLP with 32-64-128-256-256 channels), which has a slightly lower performance in most metrics. We therefore chose to use sinusoidal encodings, which also reduces the number of learnable weights.

D. Network Architecture

KPConv backbone. We show the detailed network architecture of our KPConv [45] backbone in Fig. 10. We use the same KPConv backbone as Predator, but we modify it to apply instance normalization on each point cloud individually instead of over all point clouds to allow for correct behavior over batch sizes larger than one. We do not make any other changes to the backbone for the 3DMatch dataset. However, to maintain a reasonable resolution for the downsampled keypoints for ModelNet, we use a shallower backbone consisting of only a single downsampling, and the voxel size used in the first level is set to 0.03 instead of 0.06.

Transformer. Figure 9 provides the detailed description of our transformer cross-encoder. Geometric features from the KPConv backbone are first projected to $d = 256$ dimensions, and then passed through $L = 6$ transformer cross-encoder layers to obtain the conditioned features $\bar{\mathbf{F}}_{\hat{X}}, \bar{\mathbf{F}}_{\hat{Y}}$, which can be used to predict the output correspondences and overlap scores via our output decoder. We use the pre-LN [54] configuration for the self-attention, cross-attention, and position-wise feed-forward networks (FFN). Positional encodings (Sec. C) are added to the queries, keys and values before every self- and cross-attention layer.

E. Detailed Registration Results for 3DMatch

We report the breakdown of the Registration Recall, Relative Rotation Error, and Relative Translation Error for each individual scene in Tab. 7. REGTR obtains the highest registration recall for three (3DMatch) and four (3DLoMatch) of the scenes, and the lowest rotation/translation errors for majority of the scenes in both settings, despite using downsampled features.

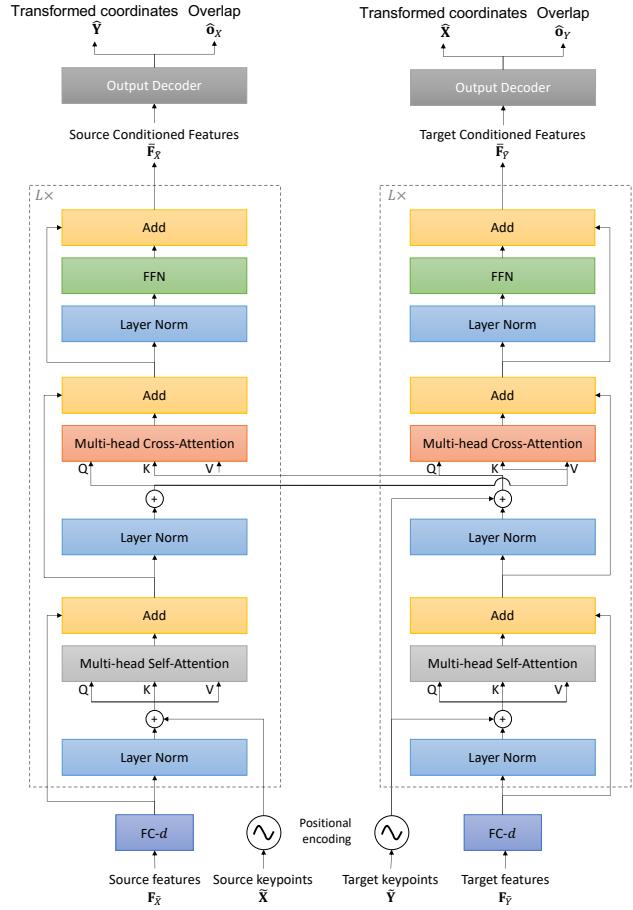


Figure 9. REGTR’s transformer cross-encoder layers.

F. Additional Qualitative Results

We show additional qualitative results for both 3DMatch and ModelNet datasets in Fig. 11. The last two rows show example failure cases. During failures, usually both overlap and correspondences are predicted wrongly.

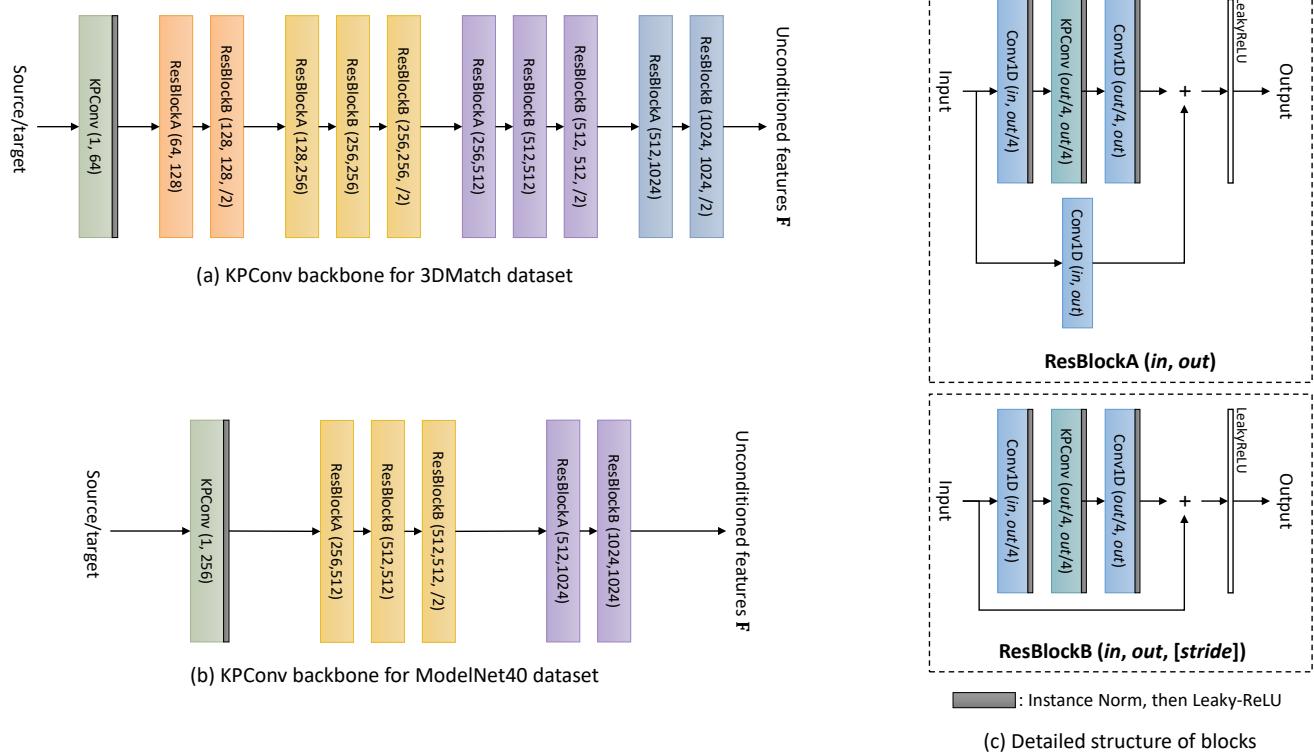


Figure 10. KPConv backbone used for (a) 3DMatch and (b) ModelNet. (c) shows the detailed structure of the residual blocks.

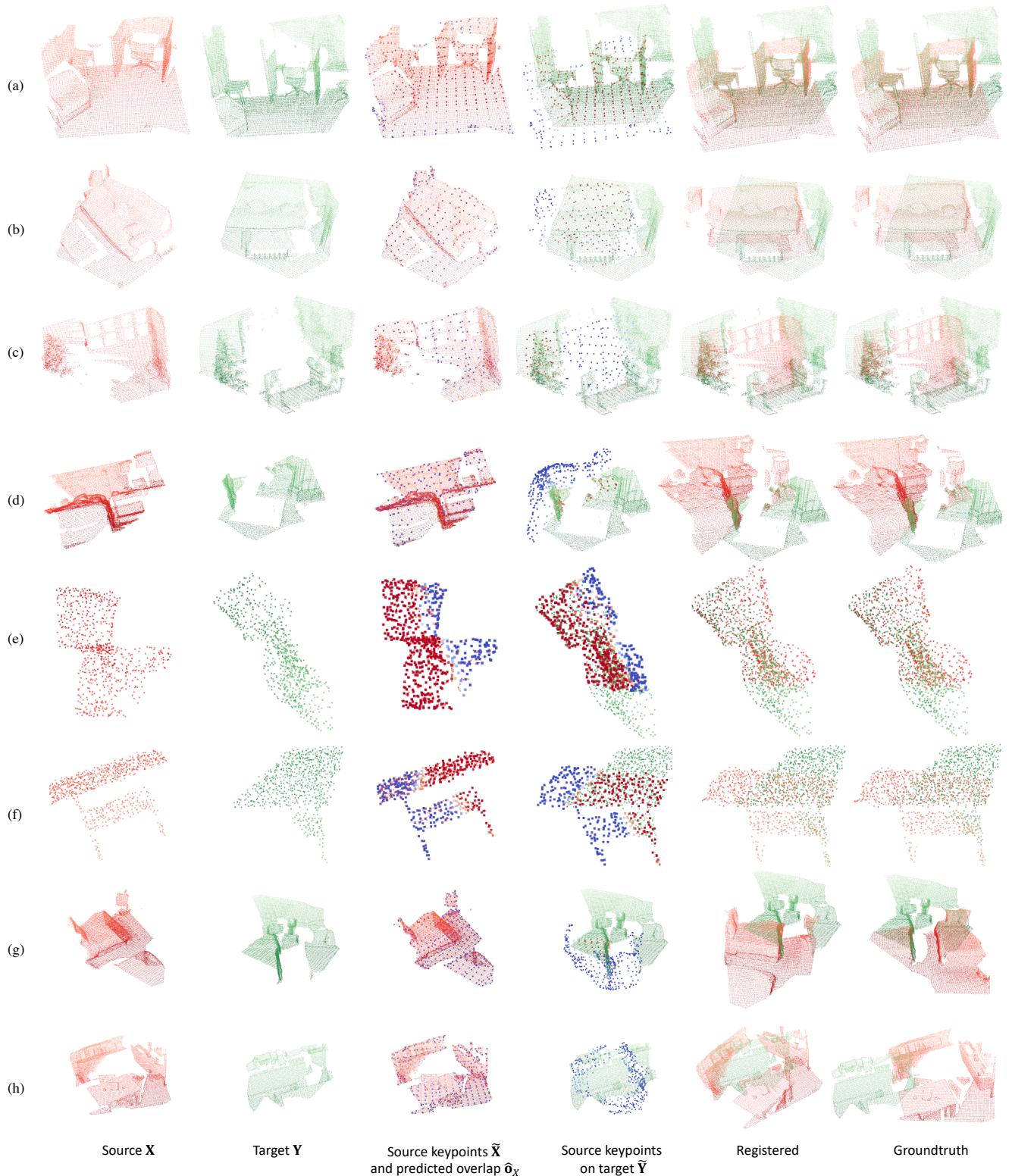


Figure 11. Additional qualitative results on (a,b) 3DMatch, (c,d) 3DLoMatch, (e) ModelNet, and (f) ModelLoNet benchmarks. Keypoints are colored by their predicted overlap scores where red indicates high overlap. The last two rows (g,h) show example failure cases on the 3DMatch dataset. Best viewed in color.

	3DMatch (\geq 30% overlap)										3DLoMatch (10-30% overlap)																													
	Kitchen					Home 1					Home 2					Hotel 1					Hotel 2					Hotel 3		Study		MIT		Lab		Avg.						
	# pairs										Registration Recall (%) \uparrow										Relative Rotation Error ($^{\circ}$) \downarrow										Relative Translation Error (m) \downarrow									
	449	106	159	182	78	26	234	45	160	524	283	222	210	138	42	237	70	191																						
3DSN [21]	90.6	90.6	65.4	89.6	82.1	80.8	68.4	60.0	78.4	51.4	25.9	44.1	41.1	30.7	36.6	14.0	20.3	33.0																						
FCGF [14]	98.0	94.3	68.6	96.7	91.0	84.6	76.1	71.1	85.1	60.8	42.2	53.6	53.1	38.0	26.8	16.1	30.4	40.1																						
D3Feat [2]	96.0	86.8	67.3	90.7	88.5	80.8	78.2	64.4	81.6	49.7	37.2	47.3	47.8	36.5	31.7	15.7	31.9	37.2																						
Predator-5k [23]	97.6	<u>97.2</u>	74.8	98.9	96.2	<u>88.5</u>	85.9	73.3	89.0	71.5	58.2	60.8	<u>77.5</u>	64.2	<u>61.0</u>	45.8	39.1	59.8																						
Predator-1k [23]	97.1	98.1	74.8	<u>97.8</u>	96.2	<u>88.5</u>	87.2	84.4	<u>90.5</u>	70.6	62.8	<u>63.1</u>	80.9	64.2	<u>61.0</u>	<u>50.0</u>	47.8	<u>62.5</u>																						
Predator-NR [23]	60.8	74.5	52.2	80.8	65.4	57.7	54.3	55.6	62.7	25.6	19.9	37.8	32.5	22.6	24.4	11.9	17.4	24.0																						
OMNet [55]	39.0	39.6	27.7	30.8	38.5	46.2	21.4	44.4	35.9	9.0	6.7	9.0	3.3	8.8	14.6	2.5	13.0	8.4																						
DGR [13]	<u>97.8</u>	94.3	62.3	95.1	88.5	84.6	84.2	75.6	85.3	60.2	45.0	52.7	54.5	48.9	41.5	38.6	<u>47.8</u>	48.7																						
PCAM [6]	96.9	93.4	78.6	96.7	83.3	84.6	81.6	68.9	85.5	<u>71.1</u>	56.7	60.8	70.8	59.9	41.5	36.4	42.0	54.9																						
Ours	97.8	90.6	<u>75.5</u>	<u>97.8</u>	<u>94.9</u>	100	88.5	91.1	92.0	66.2	<u>58.5</u>	64.9	72.7	<u>61.3</u>	70.7	53.0	71.0	64.8																						
3DSN [21]	1.926	1.843	2.324	2.041	1.952	2.908	2.296	2.301	2.199	<u>3.020</u>	3.898	3.427	3.196	3.217	3.328	4.325	3.814	3.528																						
FCGF [14]	<u>1.767</u>	1.849	2.210	1.867	1.667	2.417	2.024	1.792	1.949	2.904	3.229	3.277	2.768	2.801	2.822	3.372	4.006	3.147																						
D3Feat [2]	2.016	2.029	2.425	1.990	1.967	2.400	2.346	2.115	2.161	3.226	3.492	3.373	3.330	3.165	2.972	3.708	3.619	3.361																						
Predator-5k [23]	1.861	1.806	2.473	2.045	1.600	2.458	2.067	1.926	2.029	3.079	<u>2.637</u>	3.220	2.694	2.907	3.390	<u>3.046</u>	3.412	<u>3.048</u>																						
Predator-1k [23]	1.902	1.739	2.306	1.897	1.817	2.289	2.278	2.271	2.062	3.049	2.679	3.247	2.857	<u>2.782</u>	3.340	3.778	3.538	3.159																						
Predator-NR [23]	3.052	2.223	2.805	2.996	1.900	2.117	3.711	1.854	2.582	6.445	4.508	5.272	5.224	4.889	6.975	8.457	5.320	5.886																						
OMNet [55]	4.142	3.166	3.664	5.450	3.952	5.518	4.142	3.296	4.166	6.924	8.747	8.199	8.590	10.901	7.613	4.297	<u>3.123</u>	7.299																						
DGR [13]	2.181	1.809	2.474	<u>1.842</u>	1.966	2.313	2.653	1.588	2.103	4.049	3.967	4.433	3.666	4.119	3.742	4.188	<u>3.469</u>	3.954																						
PCAM [6]	1.965	<u>1.644</u>	<u>2.145</u>	1.874	1.434	1.631	2.250	<u>1.521</u>	<u>1.808</u>	3.501	3.518	3.571	3.649	3.197	3.278	4.148	3.368	3.529																						
Ours	1.729	1.347	1.797	1.639	1.289	<u>1.810</u>	1.570	1.357	1.567	3.366	2.446	<u>3.244</u>	<u>2.732</u>	2.439	<u>2.919</u>	3.044	2.428	2.827																						
3DSN [21]	0.059	0.070	0.079	0.065	0.074	0.062	0.093	0.065	0.071	0.082	0.098	0.096	0.101	<u>0.080</u>	0.089	0.158	0.120	0.103																						
FCGF [14]	0.053	0.056	0.071	<u>0.062</u>	0.061	0.055	0.082	0.090	0.066	0.084	0.097	<u>0.076</u>	0.101	0.084	0.077	0.144	0.140	0.100																						
D3Feat [2]	0.055	0.065	0.080	0.064	0.078	0.049	0.083	0.064	0.067	0.088	0.101	0.086	0.099	0.092	0.075	0.146	0.135	0.103																						
Predator-5k [23]	<u>0.048</u>	0.055	0.070	0.073	0.060	0.065	<u>0.080</u>	0.063	0.064	0.081	<u>0.080</u>	0.084	0.099	0.096	0.077	<u>0.101</u>	0.130	<u>0.093</u>																						
Predator-1k [23]	0.052	0.062	0.071	<u>0.062</u>	0.058	0.055	0.088	0.094	0.068	0.077	0.084	0.074	0.090	0.093	0.096	0.126	0.128	0.096																						
Predator-NR [23]	0.089	0.065	0.072	0.084	0.061	0.028	0.124	0.074	0.075	0.137	0.106	0.118	0.158	0.112	0.152	0.206	0.193	0.148																						
OMNet [55]	0.103	0.098	0.097	0.144	0.099	0.079	0.124	0.095	0.105	0.137	0.192	0.146	0.228	0.198	0.098	0.119	<u>0.094</u>	0.151																						
DGR [13]	0.057	0.064	0.070	0.068	0.052	0.062	0.094	0.072	0.067	0.104	0.111	0.119	0.111	0.092	0.085	0.147	0.134	0.113																						
PCAM [6]	0.050	<u>0.051</u>	<u>0.060</u>	0.066	<u>0.051</u>	0.058	0.081	0.052	<u>0.059</u>	0.088	0.108	0.090	0.112	0.098	<u>0.068</u>	0.117	0.110	0.099																						
Ours	0.040	0.041	0.057	0.057	0.042	<u>0.039</u>	0.054	<u>0.058</u>	0.049	<u>0.079</u>	0.064	0.078	<u>0.094</u>	0.074	0.060	<u>0.093</u>	0.077	0.077																						

Table 7. Detailed results on the 3DMatch and 3DLoMatch datasets. Results of 3DSN, FCGF, D3Feat and Predator-5k are taken from [23].