Partial Point Cloud Registration with Multi-view 2D Image Learning

Yue Zhang, Yue Wu*, Wenping Ma*, Maoguo Gong, Hao Li, Biao Hou

Xidian University, China
yzhang9812@outlook.com, {ywu@, wpma@mail., haoli@}xidian.edu.cn, gong@ieee.org, avcodec@163.com

Abstract

Learning representations from numerous 2D image data has shown promising performance, yet very few works apply this representations to point cloud registration. In this paper, we explore how to leverage the 2D information to assist the point cloud registration, and propose **IAPReg**, an **I**mage-**A**ssisted **P**artial 3D point cloud **Reg**istration framework with the multi-view images generated by the input point cloud. It is expected to enrich 3D information with 2D knowledge, and leverage 2D knowledge to assist with point cloud registration. Specifically, we create multi-view depth maps by projecting the input point cloud from several specific views, and then extract 2D and 3D features using some well-established models. To fuse the information learned from 2D and 3D modalities, inter-modality multi-view learning module is proposed to enhance geometric information and complement semantic information. Weighted SVD is a common method to reduce the impact of inaccurate correspondences on registration. However, determining the correspondence weights is not trivial. Therefore, we design a 2D-weighted SVD method, where the 2D knowledge is employed to provide weight information of correspondences. Extensive experiments perform that our method outperform the state-of-the-art method without additional 2D training data.

Introduction

Point cloud registration is a key concept in the field of 3D vision, and has been extensively used in a variety of vision tasks, including 3D object detection (Wang et al. 2023a; Hu, Liu, and Hu 2023), scene reconstruction (Xu et al. 2023; Sayed et al. 2022) and others (Yuan et al. 2023b, 2024b,a; Wu et al. 2022). The goal of point cloud registration is to find the rotation and translation parameters that align a pair of point clouds. In the real scenario, occlusions from surrounding objects and different viewing angles frequently result in incomplete point clouds acquired by the 3D scanner. These outliers, or the points in the other point cloud that have no correspondences, make point cloud registration even more difficult.

Although there are several registration methods (Huang et al. 2021; Fu et al. 2021) that can alleviate the problem of partial registration, the establishment of the correspondence is still challenging due to the limitation of geometric information and imprecise correspondences. 2D images, as a significant way to depict the 3D world, enable people to understand the 3D environment easily. At the same time, image and point cloud may be acquired by the different sensors, and the emphasis placed on them in their representations vary. This indicates that complementing information may be carried by the two modalities and fusing them effectively may be the source of unlocking the better registration performance. Nevertheless, although the fact that 2D-to-3D learning has started to receive attentions (Chen et al. 2023; Zhang et al. 2023a), the literature on the topic of partial point cloud registration still largely focuses on the single-modality (Mei et al. 2023a; Zhang et al. 2022b). Therefore, we ask the question: *how to leverage the 2D knowledge to boost partial point cloud registration?*

To tackle this issue, we present **IAPReg**, an **I**mage-**A**ssisted **P**artial point cloud **Reg**istration framework that leverages specific knowledge in 2D modality to assist point cloud registration. As shown in Figure 1, with the assistance of 2D information, our IAPReg can learn the 2D-3D joint feature to produce high-quality matching map, and then the correspondence confidence generated by 2D knowledge is used as the weight of the Singular Value Decomposition (SVD) to estimate rigid transformation.

Specifically, different from existing methods (Zhang et al. 2022a; Wang et al. 2022b) to introduce the extra RGB training images, we project the point cloud from a few specific viewpoints into multi-view depth maps since the majority of 3D sensors do not directly collect 2D images corresponding to the point clouds. To bridge the gap between 2D and 3D modalities, the inter-modality multi-view learning module consumes the features extracted from the 2D and 3D encoders, and outputs the 2D-3D joint feature. There are inevitable outliers in point clouds, and they could interfere with the establishment of correspondence. Weighted SVD method is commonly used to reduce the interference but determining the correspondence weights is not trivial. Therefore, we propose 2D-weighted SVD method to assist in calculating the weights by the well-established 2D model and estimate the rigid transformation. We conduct experiments on extensive benchmark datasets, and the experimental results demonstrate the performance of our method.

*Corresponding author.

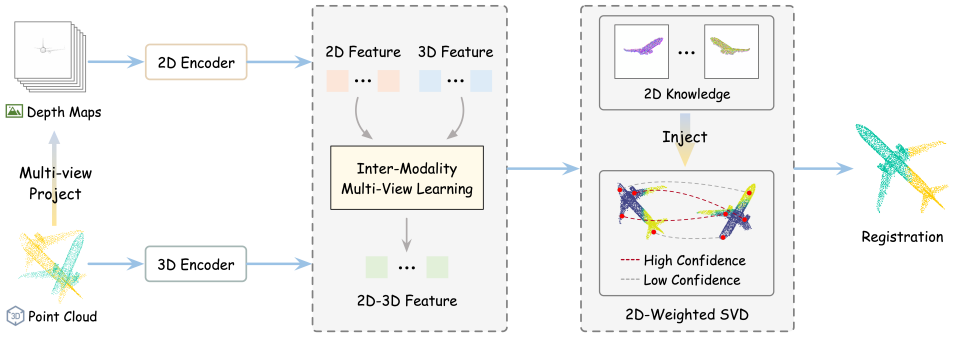


Figure 1: Our IAPReg. We combine 2D feature and 3D feature to obtain the richer representations, and leverage the 2D knowledge to guide the SVD algorithm.

Our contributions are summarized as follows:

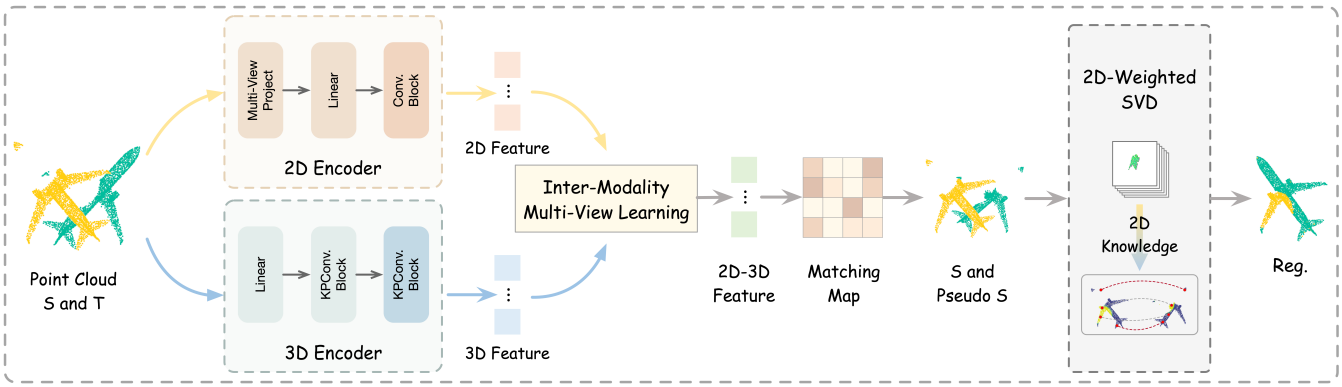
- We propose an Image-Assisted Partial Registration framework (IAPReg) for 3D point cloud registration, where 2D specific knowledge is used to enforce 3D features and assist registration.
- We propose a inter-modality multi-view learning module to enrich the 3D knowledge, and a 2D-weighted SVD method that utilizes 2D knowledge to calculate the correspondence confidence.
- We conduct extensive experiments on partial-to-partial synthetic and real-world datasets for registration demonstrating that our method achieves state-of-the-art performance without providing additional image data.

Related Work

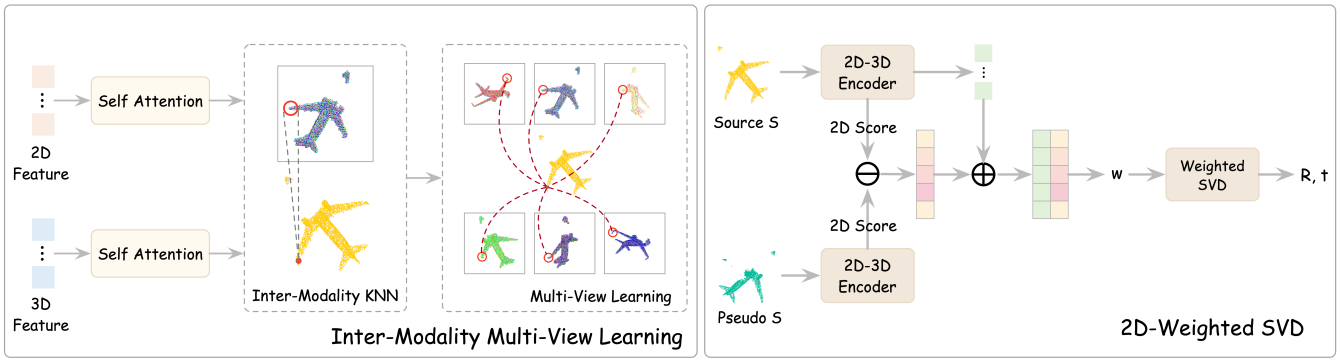
Point Cloud Registration. Point cloud registration aims to yield a 3D rigid transformation, aligning the source point cloud to the target. Iterative Closest Point (ICP) (Besl and McKay 1992) is arguably the most well-known and widely used point cloud registration algorithm, yet is prone to falling into locally optimal solutions and failing to solve the problem of partial overlap. Some methods following the ICP such as GO-ICP (Yang et al. 2015) and Super-4PCS (Mellado, Aiger, and Mitra 2014), may not provide an ideal solution under the poor initial conditions. Inspired by ICP, DCP (Wang and Solomon 2019) proposes a learning-based network extracting point cloud deep features to replace coordinate features in ICP. MAC (Zhang et al. 2023b) presents a 3D registration method with maximal cliques to mine additional local consensus knowledge in a graph for pose estimation. In order to prevent their mutual influence from degrading performance, DetarNet (Chen, Yang, and Tao 2022) proposes a Siamese Network based registration framework to decouple the translation and rotation. GraphSCNet (Qin et al. 2023) devises a non-rigid correspondence encoder based on attention to learn the robust representation for non-rigid point cloud registration. Generally, these methods only work better when the two point clouds overlap completely or have a high overlap rate, and their performance performs poorly significantly as the overlap rate decreases.

Partial Overlap Registration. In the real-world scenario, registration of partial overlap point clouds is presented as a more practical topic. MaskNet (Sarode et al. 2020) proposes a mask-based network to perform inliers estimation to improve partial registration. VBReg (Jiang et al. 2023) develops a framework based on novel variational non-local network to reformulate the non-local feature learning with variational Bayesian inference for outlier removal. To alleviate the feature distinctiveness degradation problem caused by non-overlapping points, PEAL (Yu et al. 2023b) incorporates prior knowledge into the learning process to detect the overlap region. In order to find correspondences, PCAM (Cao et al. 2021) presents a cross-attention point-wise product that allows combining both high-level contextual information and low-level geometric information. PREDATOR (Huang et al. 2021) proposes an overlap attention module,which allows information flow between the latent features of a pair of point clouds in order to address the low overlap issue. UDPReg (Mei et al. 2023b) proposes an unsupervised deep probabilistic registration framework, which applies the Sinkhorn algorithm to estimate the correspondences at the distribution level with the limitation of the mixing weights of Gaussian mixture models. To learn geometric feature for robust superpoint matching, GeoTransformer (Qin et al. 2022) proposes an encoder that embeds pair-wise distances and triplet-wise angles for partial registration. Recently proposed STORM (Wang et al. 2023b) presents an overlap matching method based on structure for partial point cloud registration, facilitating exact partial correspondence generation. Unlike these methods, our approach combines the knowledge of 2D and 3D modalities for matching map construction, and further obtains the correspondence confidence with the 2D knowledge for transformation estimation.

Image-to-Point Cloud Learning. Researchers started paying particular attention to the 2D image and 3D point cloud joint learning as a result of success in the 2D field. I2P-MAE (Zhang et al. 2023a) uses the 2D information that has been learned to drive 3D masked autoencoding, which uses an encoder-decoder architecture to reconstruct the masked point tokens. In order to enhance the cross-modal synergy found in previous studies, PiMAE (Chen et al. 2023) presents a self-supervised pretraining framework that encourages 2D and 3D interaction. A cross-modality learning



(a) Pipeline of our IAPReg.



(b) Architecture of Sub-modules.

Figure 2: Overview of our network.

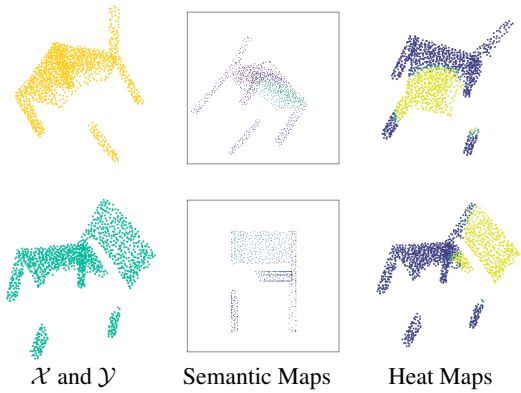


Figure 3: Visualization of semantic maps and heat maps. Darker colors indicate higher score.

approach between RGB data and point clouds is proposed by PCR-CG (Zhang et al. 2022a) to insert the deep information obtained from color signals into the geometry representation. GAVE (Wang et al. 2022b) applies extra RGB-D data to fuse the geometric and visual information. XMFNet (Aiello, Valsesia, and Magli 2022) suggests augmenting the 3D self-reconstruction losses by a differentiable renderer to test the fidelity of the completed point cloud in order to investigate how the side information might be used to shape complement. To improve 3D point cloud shape analysis, cross-modal learning and a teacher-student framework are developed by PointCMT (Yan et al. 2022) as a knowledge distillation problem. Our approach aims to assist the partial point cloud registration with the specific knowledge in 2D modality and without the extra image training data, and only leverage the multi-view images generated by the input point cloud to assist transformation estimation.

Method

In this section, we demonstrate the proposed IAPReg for partial-to-partial point cloud registration. We provide a brief introduction of our framework first and then the components are described in the following subsections.

Problem Formulation and Preliminaries

Estimating a rigid transformation $\{\mathbf{R}, \mathbf{t}\}$ to align a source point cloud $\mathcal{X} = \{\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_m\} \in \mathbb{R}^{m\times 3}$ and a target point cloud $\mathcal{Y} = \{\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_n\} \in \mathbb{R}^{n\times 3}$ is known as rigid point cloud registration, where the translation vector is denoted by $\mathbf{t} \in \mathbb{R}^3$ and the rotation matrix by $\mathbf{R} \in \mathrm{SO}(3)$, respectively. In $\mathcal{X}$ and $\mathcal{Y}$, the subscripts m and n denote the number of points, respectively. For a pair of partial-to-partial point clouds $\mathcal{X}$ and $\mathcal{Y}$, the goal is to find the overlap points

Model	ModelNet			ModelLoNet		
	RR(%) ↑	RRE(°) ↓	RTE(m) ↓	RR(%) ↑	RRE(°) ↓	RTE(m) ↓
ICP (⋆) (Besl and McKay 1992)	10.75	13.74	0.132	8.58	24.13	0.224
DCP-v2 (⋆) (Wang and Solomon 2019)	15.28	11.98	0.171	11.33	16.50	0.300
OMNet (⋄) (Xu et al. 2021)	78.67	2.947	0.032	80.10	5.223	0.118
FINet (⋄) (Xu et al. 2022)	83.82	2.712	0.039	83.56	4.835	0.103
RPMNet (⋄) (Yew and Lee 2020)	93.12	1.712	0.018	75.30	7.342	0.124
RIENet (⋄) (Shen et al. 2022)	88.75	2.447	0.018	13.23	14.49	0.105
Predator-50k(⋄) (Huang et al. 2021)	93.58	1.739	0.019	80.22	5.235	0.132
RegFormer (⋄) (Liu et al. 2023)	13.37	11.428	0.429	11.30	18.66	0.624
REGTR (⋄) (Yew and Lee 2022)	95.37	1.473	0.014	84.87	3.930	0.087
GeoTransformer (⋄) (Qin et al. 2022)	90.33	2.145	0.020	83.22	4.741	0.103
UDPReg (⋄) (Mei et al. 2023b)	<u>95.72</u>	<u>1.331</u>	<u>0.011</u>	<u>86.38</u>	<u>3.578</u>	<u>0.069</u>
Ours(⋄)	**96.58**	**1.138**	**0.010**	**89.33**	**2.842**	**0.031**

Table 1: The partial-to-partial registration results on *ModelNet40* and *ModelLoNet40*. (⋆) and (⋄) denote the non-partial registration and partial registration methods, respectively. **Boldfaced** numbers highlight the best and the second best are <u>underlined</u>.

(inliers), i.e., the correspondence of the point can be found in another point cloud. Whether the point x_i is the inlier can be determined using the following formula

$$\begin{cases} \|\mathbf{R}\mathrm{x}_i + \mathbf{t} - \mathcal{M}_{\mathcal{Y}}(\mathrm{x}_i)\|_2 \leq \zeta, & \text{inlier} \\ \|\mathbf{R}\mathrm{x}_i + \mathbf{t} - \mathcal{M}_{\mathcal{Y}}(\mathrm{x}_i)\|_2 > \zeta, & \text{outlier} \end{cases} \tag{1}$$

where the $\|\cdot\|_2$ represents the 2-Norm, the mapping function $\mathcal{M}_{\mathcal{Y}}(\mathrm{x}_i)$ outputs the corresponding point of x_i in point cloud $\mathcal{Y}$, and ζ is a tolerance that depends on the point density and distance.

The mapping function $\mathcal{M}_{\mathcal{Y}}(\mathrm{x}_i)$ can be obtained by the method based on minimum distance or soft correspondence method. In other words, it is simple to estimate the rigid transformation once the overlap points are located. Therefore, the primary challenge is to identify trustworthy overlap points that allow two point clouds can overlap as complete as possible.

Overview

As shown in Figure 2, IAPReg consists of three components: 2D encoder, 3D encoder, inter-modality multi-view learning module and 2D-weighted SVD module. Given the point clouds $\mathcal{X}$ and $\mathcal{Y}$, we first utilize a 2D encoder to embed the several depth maps projected by the input point cloud and a 3D encoder to obtain the geometrical feature. To promote the combination of 2D information and 3D information, the inter-modality multi-view learning module takes 2D and 3D features, and produces the 2D-3D joint feature. Benefiting from the 2D-3D feature, the matching map of the $\mathcal{X}$ and the $\mathcal{Y}$ can express the point-wise distance more accurately. To further relieve the affect of outliers remained in overlap estimation, we design the 2D-weighted SVD module, which further leverages 2D knowledge to calculate the correspondence confidence. In this way, the error in overlap estimation can be reduced to estimate transformation matrix accurately.

2D-3D Encoder

To compensate for the singularity and limitation of 3D information, we leverage the 2D model (such as ResNet (He et al. 2016), ViT (Dosovitskiy et al. 2020)) to assist the point cloud information extraction. Without loss of generality, the point cloud $\mathcal{X}$ serves as our example, and it is identical to the point cloud $\mathcal{Y}$.

Multi-view Projection. Considering the case of missing 2D data corresponding to the point cloud, we generate 2D data from the input point cloud. Specifically, we project the input point cloud from six viewpoints (front, back, left, right, top and bottom) and the depth information is repeated three times to generate 2D depth maps $\{\mathbf{I}_l\}_{l=1}^{6}$, where $\mathbf{I}_l \in \mathbb{R}^{H\times W\times 3}$. For the each point, we ignore the coordinates of the third dimension, and project the remaining two coordinates into the map to obtain the pixel locations. The third dimension of the point are set to pixel values to reflect depth information. Since systematic errors in the sensor could result in holes in the raw depth map, we fill the holes using the method in LLT (Wang et al. 2022b).

2D Encoder. Then, each depth map $\mathbf{I}_l$ is encoded using the pretrained ViT to obtain 2D visual feature maps $\mathbf{F}_v^l \in \mathbb{R}^{H_1\times W_1\times C_1}$. The calculation formula is as follows

$$\mathbf{F}_v^l = \mathrm{E}_v(\mathbf{I}_l) \tag{2}$$

where $\mathrm{E}_v(\cdot)$ represents the visual encoder. Such 2D visual features are more inclined to carry the semantic and texture information, and the missing geometric information can be replenished by the 3D encoder.

3D Encoder. The geometric information in 3D space is equally crucial, and we employ the KPConv (Thomas et al. 2019) as the 3D encoder to obtain the 3D geometric feature $\mathbf{F}_g \in \mathbb{R}^{m\times c}$, the calculation formula is as follows

$$\mathbf{F}_g = \mathrm{E}_g(\mathcal{X}) \tag{3}$$

where $\mathrm{E}_g(\cdot)$ represents the geometric encoder.

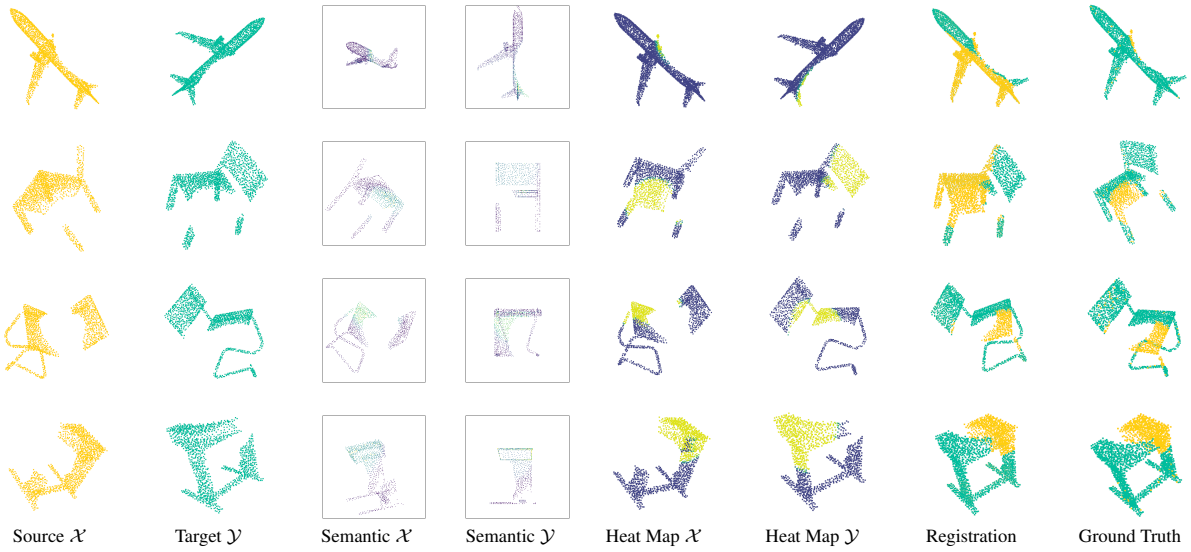


Figure 4: Qualitative registration results on ModelNet. **Heat Map:** Darker colors indicate higher weight.

Inter-Modality Multi-View Learning. The 2D and 3D features carry the specific knowledge of their modalities, and the effective fusion of knowledge could contain more abundant information and improve the subsequent tasks. Most of the previous methods (Zhang et al. 2022a; Huang et al. 2022) adopt Multilayer Perceptron (MLP) or cross-modal attention to fully autonomously learn and fuse information and ignore the prior knowledge, which may enhance redundant and non-discriminative information.

Therefore, we propose Inter-Modality Multi-View Learning module to obtain the prior inter-modal neighborhood information, and then fuse the pixel-point information. Given the x_i and its 3D geometric feature $\mathbf{F}_{g_i}$, we first project the the corresponding 2D visual feature map to 3D space by the back-projection method proposed by I2P-MAE (Zhang et al. 2023a). For the point x_i, we search k neighbor pixel features $\left\{\mathbf{P}^l = \{\mathbf{F}^l_{v_1}, \mathbf{F}^l_{v_2}, ..., \mathbf{F}^l_{v_k}\}\right\}^6_{l=1}$ by K-Nearest Neighbor (KNN) method of each view. Then, the neighbor pixel features $\mathbf{P}^l$ are aggregated, and concatenated with other view neighborhood features and 3D geometric feature to obtain the 2D-3D joint features $\bar{\mathbf{F}}^{\mathcal{J}} \in \mathbb{R}^{m\times c}$:

$$\hat{\mathbf{P}}^l = \mathrm{MLP}(\max(\mathbf{P}^l)) \tag{4}$$

$$\bar{\mathbf{F}}^{\mathcal{J}_i} = \mathbf{F}_{g_i} + \mathrm{MLP}(\underset{l=1,2,...,6}{\mathrm{cat}}(\mathbf{F}_{g_i}, \hat{\mathbf{P}}^l)) \tag{5}$$

where $\max(\cdot)$ and $\mathrm{cat}(\cdot)$ represent the operations of taking the maximum value and concatenation, l denotes the l-th view.

2D-Weighted SVD

The estimation of the pseudo correspondence is not exactly always correct. Weighted SVD is a common method to reduce the impact of false correspondences but the calculation of the weights is not trivial. Therefore, we propose the 2D-weighted SVD to leverage 2D knowledge to weight correspondences.

Given the joint features $\bar{\mathbf{F}}^{\mathcal{J}}_{\mathcal{X}}$ and $\bar{\mathbf{F}}^{\mathcal{J}}_{\mathcal{Y}}$ above, the pseudo corresponding point $\tilde{\mathrm{x}}_i \in \tilde{\mathcal{X}}$ of x_i is computed by the method of soft correspondence in form of

$$\tilde{\mathrm{x}}_i = \mathrm{softmax}(\mathcal{M}_i) \times \mathcal{Y}, \quad \mathcal{M}_i = [\mathcal{M}_{i,1}, \mathcal{M}_{i,2}, \ldots, \mathcal{M}_{i,n}] \tag{6}$$

$$\mathcal{M}_{i,j} = -\left\|\bar{\mathbf{F}}^{\mathcal{J}}_{\mathrm{x}_i} - \bar{\mathbf{F}}^{\mathcal{J}}_{\mathrm{y}_i}\right\|_2 \tag{7}$$

And then the semantic maps are obtained by ViT network, and are back-projected to 3D space, which are denoted as $\mathbf{S}_{\mathcal{X}} \in \mathbb{R}^{m\times 1}$ and $\mathbf{S}_{\tilde{\mathcal{X}}} \in \mathbb{R}^{m\times 1}$. Since the attention weights assigned to the class token indicate the contribution to which the points contribute to the category, the attention map of the class token at the final transformer layer is chosen as the semantic map. There is a higher confidence $\mathbf{w}_i$ in the correspondence $(\mathrm{x}_i, \tilde{\mathrm{x}}_i)$ if the contributions $\mathbf{S}_{\mathrm{x}_i}$ and $\mathbf{S}_{\tilde{\mathrm{x}}_i}$ to the category are closer. Therefore, the 2D semantic score $-(\mathbf{S}_{\mathrm{x}_i} - \mathbf{S}_{\tilde{\mathrm{x}}_i})^2$ is used as an attribute for calculating weight $\mathbf{w}_i$ in form of

$$\mathbf{w}_i = \mathrm{MLP}(\mathrm{cat}[\bar{\mathbf{F}}^{\mathcal{J}}_{\mathrm{x}_i}, -(\mathbf{S}_{\mathrm{x}_i} - \mathbf{S}_{\tilde{\mathrm{x}}_i})^2]) \tag{8}$$

where $\mathrm{cat}[\cdot,\cdot]$ means the concatenation.

Loss Function

We train our network end-to-end with the ground truth rigid transformation $\{\mathbf{R}, \mathbf{t}\}$ as supervision. Our loss consists of the following losses:

2D Depth Loss. We design the 2D depth loss, where the point cloud $\mathcal{X}$ and pseudo point cloud $\tilde{\mathcal{X}}$ are projected as depth maps and the mean absolute error is calculated for these maps,

$$\mathcal{L}_{\mathcal{I}} = \frac{1}{H\times W}\left\|\mathrm{proj}(\mathcal{T}^*(\mathcal{X})) - \mathrm{proj}(\tilde{\mathcal{X}})\right\|^2_2 \tag{9}$$

Model	2D Image	3DMatch			3DLoMatch		
		RR(%) ↑	IR(%) ↑	FMR(%) ↑	RR(%) ↑	IR(%) ↑	FMR(%) ↑
3DSN (⋄) (Gojcic et al. 2019)	-	78.4	36.0	95.0	33.0	11.4	63.6
FCGF (⋄) (Choy, Park, and Koltun 2019)	-	85.1	56.8	97.4	40.1	21.4	76.6
D3Feat (⋄) (Bai et al. 2020)	-	81.6	39.0	95.6	37.2	13.2	67.3
SpinNet (⋄) (Ao et al. 2021)	-	88.6	47.5	97.6	59.8	20.5	75.3
YOHO (⋄) (Wang et al. 2022a)	-	90.8	64.4	98.2	65.2	25.9	79.4
Predator-50k (⋄) (Huang et al. 2021)	-	89.0	58.0	96.6	59.8	26.7	78.6
CoFiNet (⋄) (Yu et al. 2021)	-	89.3	49.8	98.1	67.5	24.4	83.1
REGTR (⋄) (Yew and Lee 2022)	-	90.3	54.2	90.8	64.8	25.5	76.3
GeoTransformer (⋄) (Qin et al. 2022)	-	92.0	71.9	97.9	75.0	43.5	88.3
RoITr (⋄) (Yu et al. 2023a)	-	91.9	**82.6**	98.0	74.8	<u>54.3</u>	89.6
PCR-CG (⋄) (Zhang et al. 2022a)	RGB-D	89.4	-	97.4	66.3	-	80.4
IMFNet (⋄) (Huang et al. 2022)	RGB-D	-	-	<u>98.5</u>	-	-	80.6
PEAL (⋄) (Yu et al. 2023b)	RGB-D	**94.6**	72.4	**99.0**	**81.7**	<u>45.0</u>	**91.7**
PointMBF (⋄) (Yuan et al. 2023a)	RGB-D	93.2	72.1	96.4	75.2	43.3	87.1
Ours(⋄)	3D Projection	<u>93.8</u>	<u>72.7</u>	97.6	<u>75.9</u>	**45.2**	<u>89.6</u>

Table 2: The partial-to-partial registration results of different methods on *3DMatch* and *3DLoMatch*. **Boldfaced** numbers highlight the best and the second best are <u>underlined</u>.

where proj$(\cdot)$ represents the proposed multi-view projection and view aggregation, $\mathcal{T}^*$ denotes the application of the ground truth rigid transformation, H and W represent the height and width of the depth map, respectively.

Confidence Loss. The binary cross entropy loss is used to supervise the putative confidences, which is given by

$$\mathcal{L}_{\mathcal{S}} = -\frac{1}{m}\sum_{i}^{m} \mathbf{w}_i \log \mathbf{w}_i^* + (1-\mathbf{w}_i)\log(1-\mathbf{w}_i^*) \quad (10)$$

where $\mathbf{w}_i^*$ is the ground truth confidence being defined as

$$\mathbf{w}_i^* = \begin{cases} 1, & \|\mathcal{T}^*(\mathrm{x}_i) - \mathcal{N}(\mathcal{T}^*(\mathrm{x}_i), \mathcal{Y})\|_2 < \xi, \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

where $\mathcal{N}$ denotes the spatial nearest neighbor and ξ is a predefined distance threshold.

Correspondence Loss. We apply the mean absolute error on the x_i and pseudo corresponding point $\tilde{\mathrm{x}}_i$ in form of

$$\mathcal{L}_{\mathcal{C}} = \frac{1}{m}\sum_{i}^{m} \mathbf{w}_i^* \left\|\tilde{\mathrm{x}}_i - \mathcal{T}^*(\mathrm{x}_i)\right\|_2^2 \quad (12)$$

A weighted total of the three components $\mathcal{L} = \alpha\mathcal{L}_{\mathcal{I}} + \beta\mathcal{L}_{\mathcal{S}} + \gamma\mathcal{L}_{\mathcal{C}}$ represents our final loss.

Experiments

Implementation Details

We use the AdamW (Loshchilov and Hutter 2017) optimizer to train our network, starting with a 0.0001 learning rate and 0.0001 weight decay. For ModelNet40, we train for 200 epochs with a batch size of 4, and multiply the learning rate by 0.5 at epoch 70. For 3DMatch, we train for 70 epochs with a batch size of 4, halving the learning rate every 20 epochs. Our 2D and 3D encoder output final high-dimensional features with the dimension $c = 256$ and H and W in the multi-view projection are set to 224. In loss function, α and β are set to 0.1 and γ is set to 1.0 for all experiments. All experiments run on the AMD Ryzen 9 5950X CPU at 3.4GHz and single Nvidia RTX 3090 GPU.

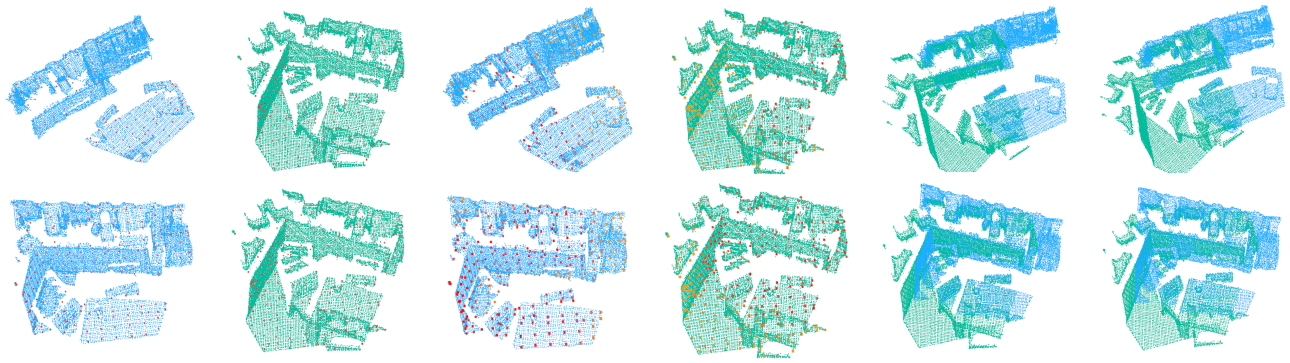
Metrics

Following (Huang et al. 2021; Qin et al. 2022), *Relative Rotation Error* (RRE) and *Relative Translation Error* (RTE) are measured the performance. In addition, We also evaluate using 1) *Inlier Ratio* (IR), the percentage of predicted correspondences under the ground-truth transformation whose residuals fall below a specific threshold (e.g., 0.1m), 2) *Feature Matching Recall* (FMR), the percentage of point cloud pairs where the inlier ratio is higher than a predetermined threshold (e.g., 30%), and (3) *Registration Recall* (RR), the percentage of pairings that properly register and whose transformation error is below a certain threshold (e.g., RMSE $<$ 0.2m).

ModelNet40

The ModelNet40 (Wu et al. 2015) consists of 12,311 meshed CAD models from 40 categories, where 9,843 models are used for training and 2,468 models are used for testing. We adopt the same data setup as (Huang et al. 2021; Yew and Lee 2022), where the point clouds are subsampled, cropped, and randomly sampled from the CAD models' mesh faces. Following (Huang et al. 2021), we use ModelNet and ModelLoNet benchmarks, which has 73.5% and 53.6% overlap rate on average, respectively.

We use modelnet40 to train both our and the comparative models and Table 1 shows the results. Remarkably, Table 1 shows that our IAPReg attains new state-of-the-art (SoTA) results. To observe the effect of the confidence score more clearly, we visualize the semantic maps and heat maps in Figure 4. In addition, we show additional qualitative results in Figure 5. In both the semantic and heat maps, it is evident



$\mathcal{X}$ and Keypoints | $\mathcal{Y}$ and Keypoints | $\mathcal{X}$ Heat Map | $\mathcal{Y}$ Heat Map | Registration | Ground Truth

Figure 5: Qualitative registration results on 3DMatch. **Red:** the keypoints after KP downsampling.

that the overlapping regions are assigned significantly higher scores than the non-overlapping regions.

3DMatch

3DMatch (Zeng et al. 2017) contains 62 scenes among which 46 are used for training, 8 for validation and 8 for testing.Utilizing preprocessed data with downsampled point clouds from (Huang et al. 2021), we evaluate both pairings with >30% overlap (3DMatch) and 10-30% overlap (3DLoMatch). The input point cloud has about 16,000 points, which are downsampled by our KPConv backbone in 3D encoder.

We conduct the experiment on 3DMatch dataset, and Table 2 shows the results. Table 2 reports that our IAPReg exhibits the second best performance, and achieves the SoTA performance compared to the methods without using additional RGB-D data. We also show several examples of visualizations, as shown in Figure 6.

Encoder		Weight	Result		
2D	3D	2D Score	RR(%) ↑	RRE(°) ↓	RTE(m) ↓
	✓		93.03	2.052	0.073
	✓	✓	94.67	1.669	0.039
✓	✓		95.21	1.329	0.040
✓	✓	✓	**96.58**	**1.138**	**0.010**

Table 3: Ablation studies on 2D knowledge.

Ablation Studies

2D Encoder. We evaluated the performance improvement of the 2D encoder. Table 3 reports that the 2D encoder achieves improvements of about 2 percentage points in *Registration Recall* and a better RTE and RRE (rows 1, 3 and 2, 4).

2D Semantic Score. We evaluate how much the 2D semantic score contributes (Eq. 8) and the results is reported in Table 3. 2D semantic score is removed and only the feature $\bar{\mathbf{F}}^{\mathcal{J}}_{\mathbf{x}_i}$ is employed to predict weights, where it can be seen that the *Registration Recall* is degraded by 1.6 percentage points for 3D encoder (rows 1 and 2), and degraded by 1.3 percentage points for 2D-3D encoder (rows 3 and 4).

Different 2D Data. We conduct experiments using three different 2D images: native depth maps in 3DMatch, RGB images in 3DMatch and our multi-view projected depth maps, where native depth map and RGB image are both single view. Table 4 reports that the proposed multi-view projection method outperforms native single-view depth maps and performs almost as well as the RGB images.

2D Data	RR(%) ↑	IR(%) ↑	FMR(%) ↑
Depth Map	90.2	71.1	96.4
RGB Image	**94.7**	**75.2**	97.0
Multi-view Projection	93.8	72.7	**97.6**

Table 4: Registration of different 2D data on 3DMatch.

Conclusion

In this paper, we propose IAPReg, a partial point cloud registration framework with 2D multi-view information. We introduce 2D knowledge from two aspects to assist point cloud registration. First, we design a inter-modality multi-view learning module that can enrich 3D information with 2D multi-view depth maps projected by the input point cloud. Next, we propose a 2D-weighted SVD module, where 2D information is used to further provide a source of confidences for correspondences. Finally, we expect our IAPReg can inspire more works to further combine multi-modality learning with point cloud registration.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62276199, 62276200, 62036006, 62171347), the Joint Funds of the National Natural Science Foundation of China (U22B2054), the Key Scientific Technological Innovation Research Project of the Ministry of Education, the China Postdoctoral Science Foundation funded project (2019M663634), the China Postdoctoral Science Foundation Special funded project (2020T130492).

References

Aiello, E.; Valsesia, D.; and Magli, E. 2022. Cross-modal Learning for Image-Guided Point Cloud Shape Completion. *Advances in Neural Information Processing Systems*, 35: 37349–37362.

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

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

Besl, P. J.; and McKay, N. D. 1992. Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, 586–606.

Cao, A.-Q.; Puy, G.; Boulch, A.; and Marlet, R. 2021. PCAM: Product of cross-attention matrices for rigid registration of point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, 13229–13238.

Chen, A.; Zhang, K.; Zhang, R.; Wang, Z.; Lu, Y.; Guo, Y.; and Zhang, S. 2023. Pimae: Point cloud and image interactive masked autoencoders for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5291–5301.

Chen, Z.; Yang, F.; and Tao, W. 2022. Detarnet: Decoupling translation and rotation by siamese network for point cloud registration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 401–409.

Choy, C.; Park, J.; and Koltun, V. 2019. Fully convolutional geometric features. In *Proceedings of the IEEE International Conference on Computer Vision*, 8958–8966.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fu, K.; Liu, S.; Luo, X.; and Wang, M. 2021. Robust point cloud registration framework based on deep graph matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8893–8902.

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, Q.; Liu, D.; and Hu, W. 2023. Density-Insensitive Unsupervised Domain Adaption on 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17556–17566.

Huang, S.; Gojcic, Z.; Usvyatsov, M.; Wieser, A.; and Schindler, K. 2021. Predator: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4267–4276.

Huang, X.; Qu, W.; Zuo, Y.; Fang, Y.; and Zhao, X. 2022. IMFNet: Interpretable multimodal fusion for point cloud registration. *IEEE Robotics and Automation Letters*, 7(4): 12323–12330.

Jiang, H.; Dang, Z.; Wei, Z.; Xie, J.; Yang, J.; and Salzmann, M. 2023. Robust Outlier Rejection for 3D Registration with Variational Bayes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1148–1157.

Liu, J.; Wang, G.; Liu, Z.; Jiang, C.; Pollefeys, M.; and Wang, H. 2023. RegFormer: an efficient projection-aware transformer network for large-scale point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8451–8460.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mei, G.; Tang, H.; Huang, X.; Wang, W.; Liu, J.; Zhang, J.; Van Gool, L.; and Wu, Q. 2023a. Unsupervised Deep Probabilistic Approach for Partial Point Cloud Registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13611–13620.

Mei, G.; Tang, H.; Huang, X.; Wang, W.; Liu, J.; Zhang, J.; Van Gool, L.; and Wu, Q. 2023b. Unsupervised Deep Probabilistic Approach for Partial Point Cloud Registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13611–13620.

Mellado, N.; Aiger, D.; and Mitra, N. J. 2014. Super 4pcs fast global pointcloud registration via smart indexing. In *Computer Graphics Forum*, volume 33, 205–215.

Qin, Z.; Yu, H.; Wang, C.; Guo, Y.; Peng, Y.; and Xu, K. 2022. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11143–11152.

Qin, Z.; Yu, H.; Wang, C.; Peng, Y.; and Xu, K. 2023. Deep graph-based spatial consistency for robust non-rigid point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5394–5403.

Sarode, V.; Dhagat, A.; Srivatsan, R. A.; Zevallos, N.; Lucey, S.; and Choset, H. 2020. MaskNet: A Fully-Convolutional Network to Estimate Inlier Points. In *Proceedings of the International Conference on 3D Vision*, 1029–1038.

Sayed, M.; Gibson, J.; Watson, J.; Prisacariu, V.; Firman, M.; and Godard, C. 2022. SimpleRecon: 3D reconstruction without 3D convolutions. In *European Conference on Computer Vision*, 1–19. Springer.

Shen, Y.; Hui, L.; Jiang, H.; Xie, J.; and Yang, J. 2022. Reliable Inlier Evaluation for Unsupervised Point Cloud Registration. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Thomas, H.; Qi, C. R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; and Guibas, L. J. 2019. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6411–6420.

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

Wang, Y.; Deng, J.; Li, Y.; Hu, J.; Liu, C.; Zhang, Y.; Ji, J.; Ouyang, W.; and Zhang, Y. 2023a. Bi-LRFusion: Bi-Directional LiDAR-Radar Fusion for 3D Dynamic Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13394–13403.

Wang, Y.; and Solomon, J. M. 2019. Deep closest point: Learning representations for point cloud registration. In *Proceedings of the IEEE International Conference on Computer Vision*, 3523–3532.

Wang, Y.; Yan, C.; Feng, Y.; Du, S.; Dai, Q.; and Gao, Y. 2023b. STORM: Structure-Based Overlap Matching for Partial Point Cloud Registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 1135–1149.

Wang, Z.; Huo, X.; Chen, Z.; Zhang, J.; Sheng, L.; and Xu, D. 2022b. Improving rgb-d point cloud registration by learning multi-scale local linear transformation. In *European Conference on Computer Vision*, 175–191. Springer.

Wu, Y.; Zhang, Y.; Fan, X.; Gong, M.; Miao, Q.; and Ma, W. 2022. INENet: Inliers estimation network with similarity learning for partial overlapping registration. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3): 1413–1426.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1912–1920.

Xu, H.; Liu, S.; Wang, G.; Liu, G.; and Zeng, B. 2021. Omnet: Learning overlapping mask for partial-to-partial point cloud registration. In *Proceedings of the IEEE International Conference on Computer Vision*, 3132–3141.

Xu, H.; Ye, N.; Liu, G.; Zeng, B.; and Liu, S. 2022. FINet: Dual Branches Feature Interaction for Partial-to-Partial Point Cloud Registration. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xu, X.; Guerrero, P.; Fisher, M.; Chaudhuri, S.; and Ritchie, D. 2023. Unsupervised 3D Shape Reconstruction by Part Retrieval and Assembly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8559–8567.

Yan, X.; Zhan, H.; Zheng, C.; Gao, J.; Zhang, R.; Cui, S.; and Li, Z. 2022. Let images give you more: Point cloud cross-modal training for shape analysis. *Advances in Neural Information Processing Systems*, 35: 32398–32411.

Yang, J.; Li, H.; Campbell, D.; and Jia, Y. 2015. Go-ICP: A globally optimal solution to 3D ICP point-set registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11): 2241–2254.

Yew, Z. J.; and Lee, G. H. 2020. Rpm-net: Robust point matching using learned features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11824–11833.

Yew, Z. J.; and Lee, G. H. 2022. REGTR: End-to-end Point Cloud Correspondences with Transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6677–6686.

Yu, H.; Li, F.; Saleh, M.; Busam, B.; and Ilic, S. 2021. Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *Advances in Neural Information Processing Systems*, 34: 23872–23884.

Yu, H.; Qin, Z.; Hou, J.; Saleh, M.; Li, D.; Busam, B.; and Ilic, S. 2023a. Rotation-invariant transformer for point cloud matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5384–5393.

Yu, J.; Ren, L.; Zhang, Y.; Zhou, W.; Lin, L.; and Dai, G. 2023b. PEAL: Prior-Embedded Explicit Attention Learning for Low-Overlap Point Cloud Registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17702–17711.

Yuan, M.; Fu, K.; Li, Z.; Meng, Y.; and Wang, M. 2023a. Pointmbf: A multi-scale bidirectional fusion network for unsupervised rgb-d point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17694–17705.

Yuan, Y.; Wu, Y.; Fan, X.; Gong, M.; Ma, W.; and Miao, Q. 2023b. EGST: Enhanced geometric structure transformer for point cloud registration. *IEEE transactions on visualization and computer graphics*.

Yuan, Y.; Wu, Y.; Fan, X.; Gong, M.; Miao, Q.; and Ma, W. 2024a. Inlier Confidence Calibration for Point Cloud Registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5312–5321.

Yuan, Y.; Wu, Y.; Gong, M.; Miao, Q.; and Qin, A. K. 2024b. One-nearest neighborhood guides inlier estimation for unsupervised point cloud registration. *IEEE Transactions on Neural Networks and Learning Systems*.

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

Zhang, R.; Wang, L.; Qiao, Y.; Gao, P.; and Li, H. 2023a. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21769–21780.

Zhang, X.; Yang, J.; Zhang, S.; and Zhang, Y. 2023b. 3D Registration with Maximal Cliques. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17745–17754.

Zhang, Y.; Yu, J.; Huang, X.; Zhou, W.; and Hou, J. 2022a. Pcr-cg: Point cloud registration via deep explicit color and geometry. In *European Conference on Computer Vision*, 443–459. Springer.

Zhang, Z.; Sun, J.; Dai, Y.; Zhou, D.; Song, X.; and He, M. 2022b. End-to-end learning the partial permutation matrix for robust 3D point cloud registration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3399–3407.