2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

# Boosting 3D Point Cloud Registration by Transferring Multi-modality Knowledge

Mingzhi Yuan[1†], Xiaoshui Huang[2†], Kexue Fu[1†], Zhihao Li[1] and Manning Wang[1✉]

***Abstract*— The recent multi-modality models have achieved great performance in many vision tasks because the extracted features contain the multi-modality knowledge. However, most of the current registration descriptors have only concentrated on local geometric structures. This paper proposes a method to boost point cloud registration accuracy by transferring the multi-modality knowledge of pre-trained multi-modality model to a new descriptor neural network. Different to the previous multi-modality methods that requires both modalities, the proposed method only requires point clouds during inference. Specifically, we propose an ensemble descriptor neural network combining pre-trained sparse convolution branch and a new point-based convolution branch. By fine-tuning on a single modality data, the proposed method achieves new state-of-the-art results on 3DMatch and competitive accuracy on 3DLoMatch and KITTI. The code and the trained model will be released at `https://github.com/phdymz/DBENet.git`.**

## I. Introduction

Multi-modality data has been demonstrated inspiring performance in numerous vision tasks. Typical examples are the recent unsupervised general models, such as CLIP [30], Flangmigo [1], which achieve accurate and high generalization performance in vision tasks. The multi-modality model usually contains plentiful of knowledge from several vision modalities. However, training such a general model requires tremendous computational and storage resources. It is usually difficult for common researchers to train such a model. Developing a multi-modality model becomes much more challenging in the 3D computer vision as the 3D multi-modality data acquisition is very expensive. In this paper, we focus on a specific task by investigating how to use the multi-modality information to solve the point cloud registration. We are not going to train such a multi-modality model. We propose a method to transfer the knowledge of an existing multi-modality model to the point cloud registration task.

3D point cloud registration [17] aims at estimating a transformation between two unaligned point clouds, which is critical to many applications including robotics [27], autonomous driving [24], and SLAM [41]. Current state-of-the-art methods [15], [35] commonly start from deep feature extraction and matching, followed by robust model fitting methods e.g. RANSAC [9] for robust transformation estimation. Although a series of distinctive deep features [7], [33], [11], [2] have been proposed recently, point cloud registration in practical scenes remains challenging due to reliance on geometric information and ignorance of semantic information.

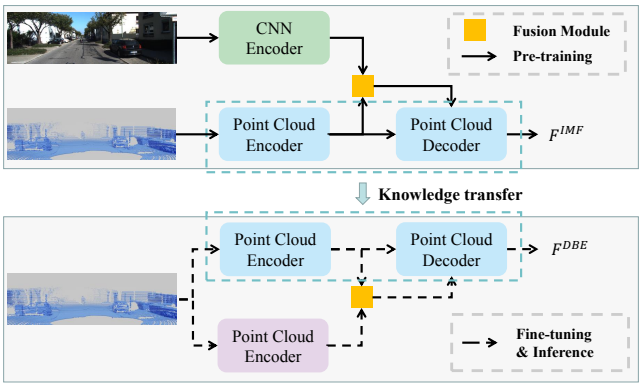


Fig. 1. **The overall architecture of our proposed method.** The network above is a dual-branch network pre-trained on multiple modalities. The network below is a dual-branch network fine-tuned on point cloud. We transfer the knowledge learned from multiple modalities to the network below by transferring the weights of pre-trained point cloud branch. Benefiting from knowledge transfer, the network below can achieve comparable performance to the above network but with only single modality input at inference. $F^{IMF}$ and $F^{DBE}$ both represent the feature matrix for point cloud registration.

In the current research stream of point cloud registration, plenty of deep descriptors [15], [35], [7], [4] only concentrate on local geometric structures. However, in practical applications, there may exist many repeatable and ambiguous structures in point clouds, such as floors, ceilings, and walls, which tend to lead to wrong matches. A feasible solution is to incorporate extra semantic information to make descriptors more reliable. Recently, IMFNet [18] introduced an additional image encoder and fused geometric information in point cloud with semantic information in image using a cross-attention. Since features generated from two different modalities have complementary information, IMFNet successfully improved the original pure point cloud descriptor by a large margin. However, it's not easy to take multi-modality data as input since IMFNet requires both RGB and point cloud during inference. Specifically, it's tedious to deal with spatial and temporal calibration and synchronization for different sensors. Moreover, acquiring multi-modality data will inevitably reduce the fault tolerance of the system, since the breakdown of either sensor can lead to failure of registration.

[1]Mingzhi Yuan, Kexue Fu, Zhihao Li and Manning Wang are with the Digital Medical Research Center, School of Basic Medical Science, Fudan University, Shanghai 200032, China, and also with the Shanghai Key Laboratory of Medical Image Computing and Computer Assisted Intervention, Shanghai, China, 200032, {mzyuan20, fukexue, mnwang}@fudan.edu.cn, lizhihao21@m.fudan.edu.cn

[2]Xiaoshui Huang is with the Shanghai artificial intelligence Lab. {huangxiaoshui}@pjlab.org.cn

† These authors contributed equally.

✉ Corresponding author.

To tackle above problem, an intuitive solution is to transfer the knowledge learned from multiple modalities to single modality. We find that a point cloud branch in multi-modality network trained on multi-modality data extracts richer information than a single point cloud network trained on pure point cloud data. Therefore, in this paper, we propose an approach for 3D point cloud registration, which transfers the knowledge of existing pre-trained multi-modality model to a new neural network that only needs point clouds during inference. As shown in Figure 1, our method utilizes two dual-branch networks. The network above is a existing pre-trained network, which takes both point cloud and image as input. The network below retains the pre-trained point cloud branch, and replace the image branch with an extra point cloud network. The reserved branch contains the knowledge learned from multiple modalities and the other modules are fine-tuned using single modality data (point cloud). During inference, the fine-tuned network only takes point clouds as input and achieves better performance by considering the complementary information in two point cloud branches.

The main contributions of our work are:

- We propose a framework transferring knowledge learned from multi-modality pre-training, which improves the performance of 3D point cloud registration while avoiding complex calibration and synchronization of sensors at inference time.
- We provide analysis of the different strategies for knowledge transfer between point cloud and image and conduct experiments to verify it.
- We propose a learnable dual-branches ensemble descriptor for point cloud registration, which consists of two mainstream feature extractors. The ablation shows that it's more powerful than a single feature extractor.
- Our method achieves new state-of-the-art results on 3DMatch [39] and shows competitive performances on 3DLoMatch [15] and Kitti [10].

## II. Related works

### A. Point cloud registration

The methods for point cloud registration can be roughly divided into two categories: correspondence-free methods and correspondence-based methods. The former mainly consists of PointnetLK [3] and its variants [16], [36], [20]. They usually extract two global features of given point clouds and directly estimate the transformation without correspondences. However, they face challenge on real-world low-overlap point clouds. The latter usually follows a pipeline of feature extraction, correspondence generation, and robust model fitting. Recently, benefiting from great progress in deep learning, many learning-based methods were proposed. FCGF [7] first utilized sparse convolution [6] to extract deep descriptors for point cloud registration, outperforming a series of hand-craft descriptors at that time. D3Feat [4] first provided a network for extracting keypoints and descriptors simultaneously. Predator [5] introduced an overlap attention to estimate overlap region, improving registration effectively. YOHO [35] designed an ensemble model to achieve SO(3)-equivariant and proposed modified RANSAC with lower complexity. Inspired by recent success in transformer [34], [8], [23], many transformer-based methods [21], [29] were proposed, taking long-dependency into consideration. However, most works inevitably tend to generate wrong correspondences facing repeatable and ambiguous structures due to reliance on geometric information. In this work, we incorporate extra information learned from multi-modality data to mitigate it.

### B. Knowledge transfer between point cloud and image

Fusion of point cloud and image [18], [13], [22], which contains complementary information, typically improves performance. FFB6D [13] designed a bidirectional network to effectively fuse features extracted from different modalities and achieved great performances in 6D pose estimation. IMFNet [18] successfully boosted registration by utilizing a cross-attention to fuse geometric information from point cloud and semantic information from image. However, all above methods need spatial and temporal calibration and synchronization for sensors, which limits their practical applications. To reserve the improvement while obtaining more flexible inference, many works attempted to transfer knowledge learned from multiple modalities to a single modality. They are roughly divided into two technical routes: network pre-training and knowledge distillation. The former commonly incorporates a network pre-trained on different modalities and utilizes knowledge in pre-trained weights during inference. For example, many monocular 3D detection methods [25], [32] leveraged depth estimation to obtain a pre-trained network and utilized it to produce pseudo-lidar to improve monocular detection. The latter achieves improvement by setting a network trained by multi-modality data as a teacher [5], [42]. For instance, S2M2-SSD [5] trained a single-modality network to learn from a multi-modality network to obtain comparable performance but took only single-modality input at inference. In this paper, we choose the former route to boost registration. We also tried using knowledge distillation but failed, more details about our attempt are illustrated in experiments.

## III. Method

Figure 2 shows the pipeline of our framework. Our framework consists of two models, IMFNet [18] and DBENet. IMFNet is an existing model pre-trained on multiple modalities, which contains knowledge we want to transfer. DBENet is our proposed **D**ual-**B**ranch **E**nsemble network, which receives the transferred knowledge. During fine-tuning, we begin with transferring the pre-trained weights of SFCN in IMFNet to the SFCN in DBENet. Then we utilize pure point clouds to fine-tune the KPFCN and attention module in DBENet. At inference, our DBENet only takes point clouds as input and can achieve high performance benefiting from both transferred knowledge and ensemble design. The details of each component are introduced in the following subsections.

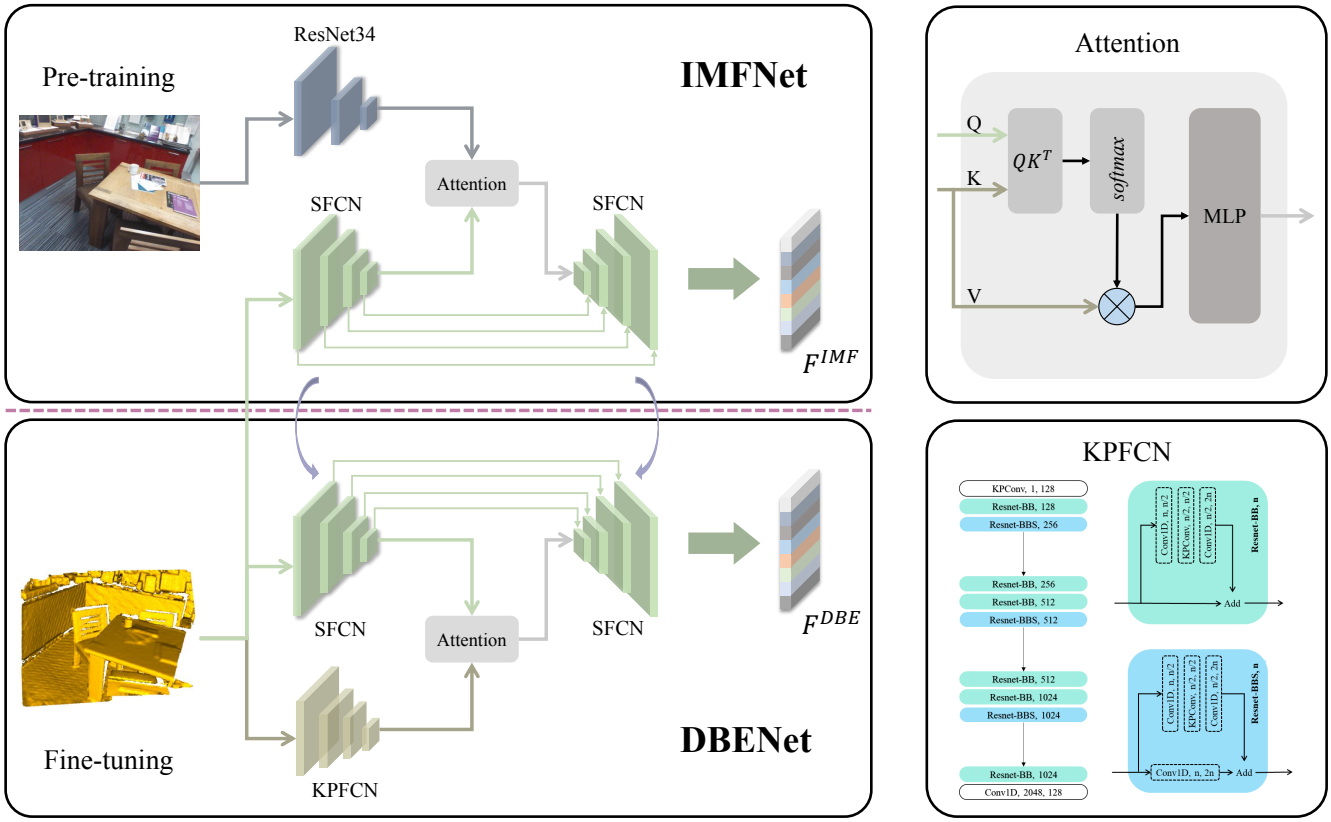


Fig. 2. **The overall architecture of our proposed framework.** Our framework contains two dual-branch networks, IMFNet and DBENet. IMFNet is an exiting model pre-trained on multiple modalities. DBENet integrates two heterogeneous branches but only takes point clouds as input. During fine-tuning, we begin with transferring the weights of SFCN in IMFNet to SFCN in DBENet. Through such weights transfer, the knowledge learned from multiple modalities can be transferred to DBENet, and then we freeze the SFCN in DBENet and fine-tune DBENet using pure point clouds. Benefiting from knowledge transfer and ensemble design for DBENet, our fine-tuned DBENet can achieve comparable or even better performance to IMFNet but with only single modality input at inference. $F^{IMF}$ and $F^{DBE}$ both represent the feature matrix for point cloud registration.

### A. IMFNet

Since IMFNet [18] has a dual-branch architecture, which separates the networks for two different modalities, we choose it as our pre-trained network. The architecture of IMFNet is shown in Figure 2, which consist of 3 components: encoder, cross-attention and decoder.

The encoder contains two branches, which encodes features from point cloud and image, respectively. The branch for point cloud i.e. sparse fully convolution network (SFCN) is implemented by sparse convolution [6] and has the same architecture as the encoder in FCGF [7].

For an input point cloud $P \in R^{N\times 3}$, it gradually abstracts it and extracts high-level features $F_P \in R^{N'\times 256}$, where $N'$ denotes the number of abstracted points. The branch for image is a ResNet34 [12]. It takes the corresponding image $I \in R^{H\times W\times 3}$ as input and outputs the flatten features $F_I \in R^{M'\times 128}$, where $M' = H/8 * W/8$ denotes the number of subsampled pixels. After obtaining features from two different modalities, IMFNet use a scale dot-product attention [34] as cross-attention module to fuse them to more reliable features $F^{IMF}_{fused} \in R^{N'\times 256}$:

$$F^{IMF}_{fused} = F_P + MLP\left(\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V\right) \quad (1)$$

where $Q = W_Q F_P \in R^{N'\times d}$ denotes the query matrix, while $K = W_K F_I \in R^{M'\times d}$ and $V = W_V F_I \in R^{M'\times d}$ denote the key matrix and the value matrix, $W_*$ denotes the learnable linear transformation mapping $F_*$ to $d$ dimension. Afterward, the fused features are gradually upsampled in U-net [31] manner to output point-wise features $F^{IMF}$ for registration.

### B. DBENet

Our DBENet also contains two different branches, but two branches take the same point clouds as input. DBENet is more powerful than a single SFCN because it intergrates a new branch based on kernel point convolution. As mentioned in [40], voxel-based methods e.g. sparse convolution has advantage of extracting coarse-grain features, while point-based methods such as PointNet++ [28], KPConv [33] has advantage of capturing fine-grain features. Therefore, we preserve the dual-branch architecture in IMFNet and replace the branch for image with a KPFCN encoder for point cloud to further improve the features extracted by sparse convolution network.

The details of KPFCN in DBENet are also shown in Figure 2. It consists of many KPConv-based blocks. Given a input point cloud $P \in R^{N\times 3}$ with features $F_{\text{in}} \in R^{N\times 1} = [1, 1 \ldots 1]^T$, KPFCN gradually encodes spatial information

into features and downsamples the point cloud. The final outputs of KPFCN are high-level features $F_P^{kp} \in R^{N' \times 128}$, which have the similar dimension with flatten image features $F_I \in R^{M' \times 128}$. We also use the same cross-attention module to fuse the features $F_P^{kp} \in R^{N' \times 128}$ extracted by KPFCN with the features $F_P \in R^{N' \times 256}$ extracted by the SFCN. After that, a decoder gradually upsamples the fuesd features $F_{fused}^{DBE} \in R^{N' \times 256}$ and normalizes the features $F^{DBE} \in R^{N \times 32}$ in the last layer as outputs for registration. We can use $F^{DBE} \in R^{N \times 32}$ to generate putative correspondences and then use a robust estimator such as RANSAC [9] to achieve registration during inference.

### C. Fine-tuning

As mentioned above, point cloud contains rich geometric information and image texture contains rich semantic information, making feature extractor pre-trained on multiple modalities more distinctive. Therefore, we first transfer the knowledge contained in pre-trained model to our DBENet.

Since our DBENet has a same SFCN with IMFNet and SFCN in IMFNet contains knowledge learned from multiple modalities, we can achieve knowledge transfer by directly transferring the network weights. Specifically, we replace the randomly initialized weights in DBENet with the pre-trained weights of SFCN in IMFNet and then freeze the transferred weights and fine-tune the weights of KPFCN and cross-attention module.

We utilize the hardest contrastive loss [7] to fine-tune our DBENet. Given a point cloud pair, we first sample some points to generate putative correspondences by feature matching and build the positive pair set $\mathcal{P}$ and the negative pair set $\mathcal{N}$. The positive pair set $\mathcal{P}$ contains feature pairs of the putative correspondences with limited residual error under the ground truth transformation, while the negative pair set $\mathcal{N}$ mines the hardest negatives $\hat{F}_i^{DBE}, \hat{F}_j^{DBE}$ for $(F_i^{DBE}, F_j^{DBE})$ in $\mathcal{P}$ and remove false negatives. The loss function is formulated as:

$$\begin{aligned}\mathcal{L} = \sum_{(i,j)\in\mathcal{P}} \{ & \left[ D\left(F_i^{DBE}, F_j^{DBE}\right) - m_p \right]_+^2 / |\mathcal{P}| \\ & + \lambda_n I_i \left[ m_n - \min_{k\in\mathcal{N}} D\left(F_i^{DBE}, F_k^{DBE}\right) \right]_+^2 / |\mathcal{P}_i| \\ & + \lambda_n I_j \left[ m_n - \min_{k\in\mathcal{N}} D\left(F_j^{DBE}, F_k^{DBE}\right) \right]_+^2 / |\mathcal{P}_j| \}\end{aligned} \tag{2}$$

where $D(\cdot)$ denotes the Euclidean distance, $I[\cdot]$ is an indicator function that return 1 if the condition is satisfied and 0 otherwise, $m_p$ and $m_n$ denote the margins for positive and negative pairs, $|\mathcal{P}_i|$ denotes the number of valid hardest negatives for the first items in $\mathcal{P}$ and $|\mathcal{P}_j|$ for the second items, $\lambda_n$ denotes the weight to balance positive loss and negative loss, we set $m_p = 0.1, m_n = 1.4, \lambda_n = 0.5$ in fine-tuning.

## IV. Experiment

We evaluate our method by comparing it with many state-of-the-art methods on widely-used public datasets including 3DMatch [39], 3DLoMatch [15], and KITTI [10]. The following sections are organized as follows. First, we illustrate our experimental settings including implementation details and evaluation metrics in section IV-A. Next, we conduct experiments on indoor datasets, 3DMatch and 3DLoMatch in section IV-B and IV-C. We also implement an experiment on outdoor dataset KITTI in section IV-D. To further understand our method, we conduct comprehensive ablation studies in section IV-E.

### A. Experimental settings

**Implementation:** We implement our networks in Pytorch [26]. For IMFNet [18], we directly use the pre-trained weights provided in `https://github.com/XiaoshuiHuang/IMFNet`. We use the ADAM optimizer [19] with an initial learning rate of 0.1 to fine-tune our DBENet for 10 epochs and the batch size for fine-tuning is set to 2. All the experiments are conducted on a single GTX 1080ti graphic card with Intel Core i7-7800X CPU.

**Evaluation metrics:** To quantitatively compare our method with other state-of-the-art methods, we select several widely-used evaluation metrics to evaluate performance.

For 3DMatch [39] and 3DLoMatch [15] benchmark, the most widely-used evaluation metrics are: Feature-Match Recall (FMR), Inlier-Ratio (IR), and Registration Recall (RR). FMR measures the percentage of pairs that have $> 5\%$ inlier correspondences with 10cm residual under the ground truth transformations. IR describes the percentage of inlier correspondences among all the putative correspondences generated by feature matching. And RR directly shows the percentage of successfully registered pairs. In these two benchmarks, we consider the registration with $E_{RMSE} < 0.2m$ as a successfully registered pair. $E_{RMSE}$ denotes the error metric between an unaligned point cloud pair $\{i, j\}$:

$$E_{RMSE} = \sqrt{\frac{1}{|\Omega^*|} \sum_{(x^*,y^*)\in\Omega^*} \left\| \hat{T}_{i,j} x^* - y^* \right\|^2} \tag{3}$$

where $\hat{T}_{i,j}$ represents the estimated transformations for point cloud pair $\{i, j\}$, $\Omega^*$ represents the set containing all the inlier correspondences, $x^*$ and $y^*$ represent the 3D coordinates in it.

For KITTI [12] benchmark, the most widely-used evaluation metrics are: Relative Translation Error (RTE), Relative Rotation Error (RRE) and Success rate (Success). RTE is defined as RTE$= \left| \hat{t} - t^* \right|$, where $\hat{t}$ denotes the estimated translation and $t^*$ denotes the ground truth translation. RRE is defined as RRE$= \arccos\left( \left( \mathrm{Tr}\left( \hat{R}^T R^* \right) - 1 \right) / 2 \right)$, where $\hat{R}$ denotes the estimated rotation and $R^*$ denotes the ground truth rotation. Success rate measures the percentage of registered pairs with RTE$< 2m$ and RRE$< 5^\circ$.

### B. Experiment on 3DMatch

3DMatch [39] is a well-known dataset, which consists of 62 scenes. Here we follow the official split [15] to divide the dataset into 46 scenes for training, 8 scenes for validation and

TABLE I

RESULTS ON 3DMATCH DATASET.

| | FMR (%) | IR (%) | RR (%) |
|---|---|---|---|
| 3DSN [11] | 94.7 | 36.0 | 78.4 |
| FCGF [7] | 95.2 | 56.8 | 85.1 |
| D3Feat [4] | 95.8 | 39.0 | 81.6 |
| Predator [15] | 96.6 | 61.0 | 88.3 |
| SpinNet [2] | 97.6 | 47.5 | 88.6 |
| YOHO [35] | 98.2 | 64.4 | 90.8 |
| CoFiNet [38] | 98.1 | 49.8 | 89.3 |
| GeoTransformer [29] | 97.9 | 71.9 | 92.0 |
| REGTR [37] | - | - | 92.0 |
| Lepard [21] | 98.3 | 55.5 | 93.5 |
| IMFNet [18] | **98.6** | 85.5 | 93.4 |
| Ours | **98.6** | **86.1** | **93.8** |

TABLE II

RESULTS ON 3DLOMATCH DATASET.

| | FMR (%) | IR (%) | RR (%) |
|---|---|---|---|
| 3DSN [11] | 63.6 | 11.4 | 33.0 |
| FCGF [7] | 76.6 | 21.4 | 40.1 |
| D3Feat [4] | 67.3 | 15.0 | 46.9 |
| Predator [15] | 78.6 | 38.0 | 56.7 |
| SpinNet [2] | 75.3 | 20.5 | 59.8 |
| YOHO [35] | 79.4 | 25.9 | 65.2 |
| CoFiNet [38] | 83.1 | 24.4 | 67.5 |
| GeoTransformer [29] | **88.3** | 43.5 | **75.0** |
| REGTR [29] | - | - | 64.8 |
| Lepard [21] | 84.5 | 26.0 | 69.0 |
| IMFNet [18] | 80.3 | 46.6 | 65.9 |
| Ours | 80.3 | **47.7** | 65.0 |

8 scenes for test. We compare our methods with many state-of-the-art methods. Besides IMFNet [18], all the competitors are trained and tested on point clouds and without refinement during test, while IMFNet is trained and tested using two different modalities.

The results of our method and competitors are shown in Table I. Our method surprisingly achieves new state-of-the-art on all evaluation metrics, and even outperforms the method using multiple modalities. Moreover, our method successfully boosts the performance of baseline i.e. FCGF [7] by a large margin. More analysis on improvement is illustrated in ablation studies.

### C. Experiment on 3DLoMatch

3DLoMatch [15] is a much more challenging benchmark, which consists of low-overlap ratio (10%-30%) point cloud pairs from 3DMatch. All the competitors tested on 3DLoMatch are previously trained on 3DMatch.

The results on 3DLoMatch are shown in Table II. It's observed that our method only achieves state-of-the-art on IR metric. This is because the state-of-the-art methods such as GeoTransformer [29] are trained using overlap-aware loss, which provides extra supervision for network to learn to estimate overlap regions. Although these overlap-based methods have a natural advantage when facing low-overlap point cloud pairs, our method still shows a competitive performance even outperforms Predator [15], which contains an overlap attention. Among methods without overlap supervision such as YOHO [35] and SpinNet [2], our method is a cost-effective choice due to its simpler implementation and comparable performance. Moreover, our method achieves comparable performance with IMFNet [18], but our method only takes point clouds as input during inference, avoiding the challenge of calibration and synchronization for sensors. As on 3DMatch benchmark, our method also boosts baseline i.e. FCGF by a large margin, indicating the effectiveness of our method.

### D. Experiment on KITTI

KITTI [10] is one of the most well-known datasets for autonomous driving, which contains 3D point clouds captured by LiDAR. We follow previous settings in [18] and divide 11 sequences (0-10) of the odometry dataset into training set (0-5), validation set (6-7) and test set (8-10). Referring to the experimental settings in [18], we report both the averages and standard deviations for RTE and RRE.

TABLE III

RESULTS ON KITTI DATASET.

| | RTE (cm) | STD (cm) | RRE (°) | STD (°) | Success (%) |
|---|---|---|---|---|---|
| FCGF [7] | 6.47 | 6.07 | **0.23** | 0.23 | 98.92 |
| D3Feat [4] | 6.90 | 0.30 | 0.24 | 0.06 | **99.81** |
| Predator [15] | 6.80 | - | 0.27 | - | 99.80 |
| SpinNet [2] | 9.88 | 0.50 | 0.47 | 0.09 | 99.10 |
| IMFNet [18] | **5.77** | **0.27** | 0.37 | **0.01** | 99.28 |
| Ours | 5.98 | 0.29 | 0.42 | **0.01** | 99.10 |

The results are shown in Table III. Although our method does not achieve the best performance, it still achieves comparable performance to the method trained on multiple modalities. Moreover, our method also boosts the performance of FCGF [7], which serves as a baseline in our experiments. This also reflects the effectiveness of our method.

### E. Ablation studies

To further understand our work, we conduct several ablation studies. First, we conduct an experiment to illustrate the effectiveness of multi-modality pre-training and our ensemble design. Second, we empirically demonstrate our decision to use the aforementioned fine-tuning strategy. Finally, we briefly discuss another knowledge transfer method, knowledge distillation and illustrate our attempt. All the ablation studies are conducted on 3DMatch dataset.

**Effectiveness of multi-modality pre-training and model ensemble.** As mentioned in previous paragraph, multi-modality data is more informative, which helps network to learn a more distinctive descriptor. We verify it by replacing the weights in an original FCGF [7] with that in IMFNet [18]. The comparison is shown in line 2 and 3 in Table IV. Here we use P to denote using the weights trained on pure point cloud, and use P+I to denote using the weights coming from pre-trained IMFNet. It's observed that the performance of FCGF using multi-modality pre-trained weights performs better, which verifies our proposal.

As illustrated in [40], voxel-based feature and point-based feature are commonly complementary. The former has advantages in learning structure, while the later has advantages in capturing details. Therefore, the ensemble model which has heterogeneous branches tends to perform better. Line 2

TABLE IV

ABLATION ON MULTI-MODALITY PRE-TRAINING AND MODEL ENSEMBLE.

| | FMR (%) | IR (%) | RR (%) |
|---|---|---|---|
| FCGF (P) | 95.2 | 56.8 | 85.1 |
| FCGF (P+I) | 97.7 | 85.4 | 92.9 |
| Ours (scratch) | 96.7 | 83.3 | 92.1 |
| Ours | **98.6** | **86.1** | **93.8** |

TABLE V

ABLATION ON FINE-TUNING STRATEGIES. ✓ DENOTES FREEZING THE PRE-TRAINED WEIGHT.

| Encoder | Attention | Decoder | FMR (%) | IR (%) | RR (%) |
|---|---|---|---|---|---|
| ✓ | | | 98.5 | 85.6 | 93.1 |
| ✓ | ✓ | | 98.3 | 85.8 | 93.1 |
| ✓ | ✓ | ✓ | 98.3 | **86.2** | 93.3 |
| ✓ | | ✓ | **98.6** | 86.1 | **93.8** |

and 4 in Table IV also verify it. We compare the baseline i.e. FCGF with a DBENet trained without using transferred weight. The DBENet trained from scratch also improves the performance of the baseline, indicating the effectiveness of our ensemble design. All in all, both the multi-modality pre-training and model ensemble help our method achieve outstanding performance on point cloud registration.

**Ablation on fine-tuning strategies.** Pre-training and then fine-tuning is a widely-used paradigm. Generally speaking, freezing encoder and fine-tuning other modules is the most conventional choice. We also attempt to freeze other pre-trained modules and find that freezing encoder and decoder and then fine-tuning attention and KPFCN is the best choice as shown in Table V.

**Discussion on knowledge transfer.** As mentioned in related works, there exists the other methods to transfer knowledge learned from multiple modalities to a single modality. Therefore, we designed a knowledge distillation strategy to achieve it.

Since point cloud is not as regular as image, the number of input points is uncertain and the extracted high-level features are unordered, making the features extracted by KPFCN impossible to approximate the features extracted by ResNet34 by a direct supervision. Therefore, we designed a point-to-point teacher-student loss in the output of the cross-attention module to avoid pixel-to-point direct distillation.

We hope the teacher i.e. pre-trained IMFNet to guides the student i.e. DBENet, so that the fused features $F^{DBE}_{fused}$ extracted by student can simulate the fused features $F^{IMF}_{fused}$ extracted by teacher. To make fused features as consistent as possible, we add a point-to-point KL loss during fine-tuning:

$$\mathcal{L}_{KD} = \frac{1}{256 * N'} \sum_{i=1}^{N} \sum_{j=1}^{256} \phi\left(F^{IMF}_{fused,i,j}\right) \cdot \log \left[ \frac{\phi\left(F^{IMF}_{fused,i,j}\right)}{\phi\left(F^{DBE}_{fused,i,j}\right)} \right] \tag{4}$$

where $\phi(\cdot)$ denotes a channel-wise softmax function [14] with temperature $T = 1$.

The results are shown in Table VI. Pre-training denotes using pre-trained weights, KD denotes using an extra point-to-point teacher-student loss during fine-tuning. Although it sounds reasonable to add a loss to make the student features simulate the teacher's as much as possible to improve student, it fails in practical experiment and even has a negative impact. We infer that the large gap between two completely different modalities and network architectures makes it fail to distillate knowledge explicitly.

TABLE VI

ABLATION ON KNOWLEDGE TRANSFER STRATEGIES.

| Pre-training | KD | FMR (%) | IR (%) | RR (%) |
|---|---|---|---|---|
| | | 96.7 | 83.3 | 92.1 |
| | ✓ | 93.1 | 78.8 | 88.2 |
| ✓ | ✓ | 98.3 | 85.8 | 93.3 |
| ✓ | | **98.6** | **86.1** | **93.8** |

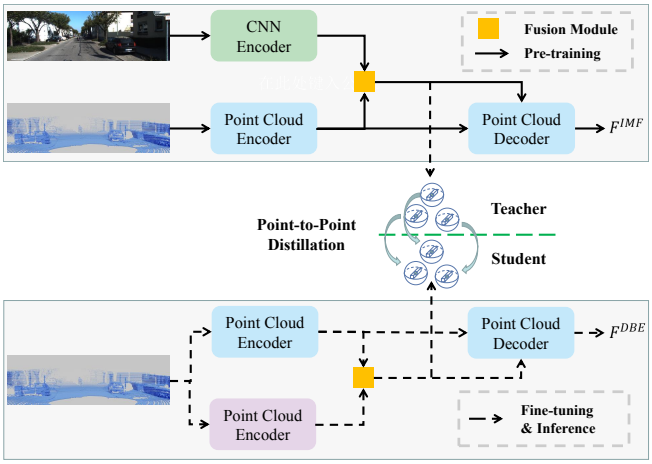


Fig. 3. Multi-modality knowledge distillation for point cloud registration.

We also attempted replacing KL loss with L1 loss, but get similar results. It can be seen that for heterogeneous cross-modality networks, pre-training may be a better way of knowledge transfer than distillation. Unified networks between different modalities such as transformer may [34], [30] have the potential for knowledge transfer in terms of knowledge distillation because it may be able to narrow the gap caused by the difference in network architectures.

## V. CONCLUSION

In this paper, we propose a method to transfer the multi-modality knowledge to boost the performance of point cloud registration. Our proposed method ensembles the pre-trained sparse convolution branch and point convolution branch, which can leverage the multi-modality knowledge and utilize only the point cloud modality during inference. The proposed method does not require the strict calibration and synchronization of multiple modalities during the inference. The experiments show that the ensemble model with multi-modality knowledge can significantly improve the registration accuracy and even outperform the multi-modality model.


## ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant 62076070.

## REFERENCES


[1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=EbMuimAbPbs

[2] S. Ao, Q. Hu, B. Yang, A. Markham, and Y. Guo, "Spinnet: Learning a general surface descriptor for 3d point cloud registration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 753–11 762.

[3] Y. Aoki, H. Goforth, R. A. Srivatsan, and S. Lucey, "Pointnetlk: Robust & efficient point cloud registration using pointnet," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7163–7172.

[4] X. Bai, Z. Luo, L. Zhou, H. Fu, L. Quan, and C.-L. Tai, "D3feat: Joint learning of dense detection and description of 3d local features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6359–6367.

[5] Z. Chong, X. Ma, H. Zhang, Y. Yue, H. Li, Z. Wang, and W. Ouyang, "Monodistill: Learning spatial features for monocular 3d object detection," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=C54V-xTWfi

[6] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3075–3084.

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

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

[9] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[10] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[11] Z. Gojcic, C. Zhou, J. D. Wegner, and A. Wieser, "The perfect match: 3d point cloud matching with smoothed densities," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5545–5554.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun, "Ffb6d: A full flow bidirectional fusion network for 6d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3003–3013.

[14] G. Hinton, O. Vinyals, J. Dean, *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[15] S. Huang, Z. Gojcic, M. Usvyatsov, A. Wieser, and K. Schindler, "Predator: Registration of 3d point clouds with low overlap," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2021, pp. 4267–4276.

[16] X. Huang, G. Mei, and J. Zhang, "Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 366–11 374.

[17] X. Huang, G. Mei, J. Zhang, and R. Abbas, "A comprehensive survey on point cloud registration," *arXiv preprint arXiv:2103.02690*, 2021.

[18] X. Huang, W. Qu, Y. Zuo, Y. Fang, and X. Zhao, "Imfnet: Interpretable multimodal fusion for point cloud registration," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 12 323–12 330, 2022.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] X. Li, J. K. Pontes, and S. Lucey, "Pointnetlk revisited," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 763–12 772.

[21] Y. Li and T. Harada, "Lepard: Learning partial point cloud matching in rigid and deformable scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5554–5564.

[22] Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le, *et al.*, "Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 182–17 191.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[24] W. Lu, G. Wan, Y. Zhou, X. Fu, P. Yuan, and S. Song, "Deepvcp: An end-to-end deep neural network for point cloud registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 12–21.

[25] X. Ma, S. Liu, Z. Xia, H. Zhang, X. Zeng, and W. Ouyang, "Rethinking pseudo-lidar representation," in *European Conference on Computer Vision*. Springer, 2020, pp. 311–327.

[26] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[27] F. Pomerleau, F. Colas, R. Siegwart, *et al.*, "A review of point cloud registration algorithms for mobile robotics," *Foundations and Trends® in Robotics*, vol. 4, no. 1, pp. 1–104, 2015.

[28] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[29] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, and K. Xu, "Geometric transformer for fast and robust point cloud registration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 143–11 152.

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[32] A. Simonelli, S. R. Bulo, L. Porzi, M. López-Antequera, and P. Kontschieder, "Disentangling monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1991–1999.

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

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

[35] H. Wang, Y. Liu, Z. Dong, W. Wang, and B. Yang, "You only hypothesize once: Point cloud registration with rotation-equivariant descriptors," *arXiv preprint arXiv:2109.00182*, 2021.

[36] H. Xu, S. Liu, G. Wang, G. Liu, and B. Zeng, "Omnet: Learning overlapping mask for partial-to-partial point cloud registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3132–3141.

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

[38] H. Yu, F. Li, M. Saleh, B. Busam, and S. Ilic, "Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 872–23 884, 2021.

[39] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser, "3dmatch: Learning local geometric descriptors from rgb-d reconstructions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1802–1811.

[40] F. Zhang, J. Fang, B. Wah, and P. Torr, "Deep fusionnet for point

cloud semantic segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 644–663.

[41] J. Zhang and S. Singh, "Visual-lidar odometry and mapping: Low-drift, robust, and fast," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 2174–2181.

[42] W. Zheng, M. Hong, L. Jiang, and C.-W. Fu, "Boosting 3d object detection by simulating multimodality on point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 638–13 647.