# PCR-CG: Point Cloud Registration via Deep Explicit Color and Geometry

Yu Zhang[1] Junle Yu[2] Xiaolin Huang[1] Wenhui Zhou[2] Ji Hou[3]

[1]Shanghai Jiaotong University [2]Hangzhou Dianzi University [3]TUM

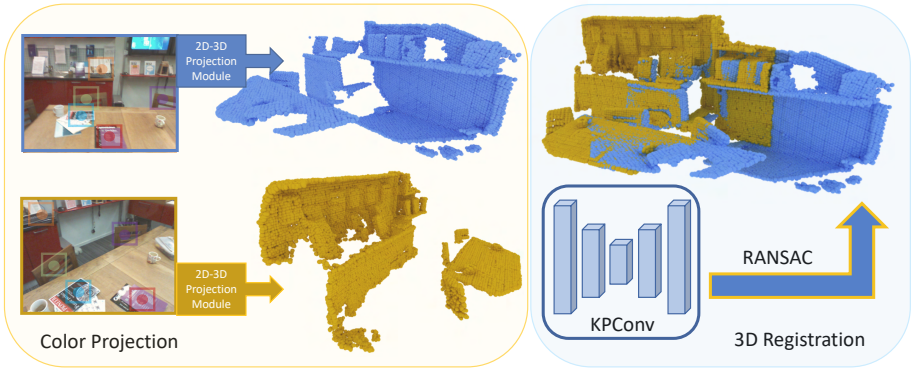


Fig. 1: We seek to align two point clouds in RGB-D data. To better leverage color, we propose PCR-CG, a 2D-3D projection module that explicitly lifts 2D deep color features to 3D geometry representation. A pair of RGB-D frames are used as input, where each RGB-D frame is composed of a color image and a depth frame. 3D geometry is represented by the point cloud that is generated from depth frame. We leverage a pre-trained 2D network to predict correspondences between frames and extract regional features from color images. The 2D regional features are further lifted to 3D via our proposed 2D-3D projection module in an explicit manner.

**Abstract.** In this paper, we introduce PCR-CG: a novel 3D point cloud registration module explicitly embedding the color signals into geometry representation. Different from the previous SOTA methods that used only geometry representation, our module is specifically designed to effectively correlate color and geometry for the point cloud registration task. Our key contribution is a 2D-3D cross-modality learning algorithm that embeds the features learned from color signals to the geometry representation. With our designed 2D-3D projection module, the pixel features in a square region centered at correspondences perceived from images are effectively correlated with point cloud representations. In this way, the overlap regions can be inferred not only from point cloud but also from the texture appearances. Adding color is non-trivial. We compare against a variety of baselines designed for adding color to 3D, such as exhaustively adding per-pixel features or RGB values in an implicit manner. We leverage Predator as our baseline method and incorporate our

module into it. Our experimental results indicate a significant improvement on the 3DLoMatch benchmark. With the help of our module, we achieve a significant improvement of 6.5% registration recall with 5000 sampled points over our baseline method. To validate the effectiveness of 2D features on 3D, we ablate different 2D pre-trained networks and show a positive correlation between the pre-trained weights and task performance. Our study reveals a significant advantage of correlating explicit deep color features to the point cloud in the registration task.

## 1 Introduction

With commodity depth sensors commonly available, such as Kinect series, a variety of RGB-D datasets are created [11,47,3,41]. With recent breakthroughs in deep learning and the increasing prominence of RGB-D data, the computer vision community has made a tremendous progress on analyzing point cloud [33] and images [18,17]. Recently, we have observed a rapid progress in cross-modality learning between geometry and colors [21,26,25,42,9,7]. However, prior work mainly focused on high-level semantic scene understanding tasks, such as semantic/instance segmentation [12,24] and object detection [31]. Compared to high-level tasks, cross-modality learning between color and geometry is less explored in low-level tasks, such as point cloud registration. In this paper, we discuss correlating RGB priors for aligning two partial point clouds.

Point cloud registration has been speedily developed because of its wide applications [23,2,46,14,5]; its 2D counter-part has been developed even earlier and achieved great success [29] in many systems, such as visual SLAM [43]. Mainstream methods adopt a first-correspondences-then-transformation manner, namely estimating transformations between two frames based on these correspondence matching. In this context, correspondence-matching-based methods [37,48,29] have showed appealing results in the 2D domain. However, current deep learning based methods in 3D merely use geometry as the only input. Therefore, exploring to combine deep RGB features is valuable and of great importance to the point cloud registration task. In this manner, a variety of existing 2D approaches and pre-trained models can also be further leveraged in 3D point cloud registration task.

Finding correspondences is essential for calculating the transformation matrix between two frames, and correspondences only appear in the overlap region. In this context, estimating the overlap region of two frames is critical for point cloud registration. Intuitively, we can identify the overlap regions not only from geometric inputs like point cloud, but also from color signals like images. Given this observation, we propose to embed color signals into point cloud representation, so as to effectively predict 3D correspondences for the registration task. To this end, we propose PCR-CG, a novel module that explicitly embeds RGB priors into the geometry representation for the point cloud registration.

In our work, we build upon the successful Predator [23], following the standard point cloud registration pipeline, namely first finding correspondences and then using RANSAC to estimate the rotation and translation matrices between

two frames of point clouds. To enable the usage of RGB values from captured RGB-D data, our approach introduces three steps. First, a 2D pre-trained neural network [37] is used to predict pixel correspondences between pure RGB frames. Based on the correspondences, we extract square regions centered at each correspondence pixel. Furthermore, the 2D pre-trained neural network summarizes the features from pixels in each region. We investigate the effectiveness of 2D pre-trained features in the 3D task by trying different 2D pre-trained weights, such as ImageNet and Pri3D [21] pre-trained models. We note that the 2D models are pre-trained on different datasets. In this context, the transfer ability of the 2D part shows promising results. In this manner, we are able to take advantage of massive existing 2D pre-trained models. Secondly, we propose a 2D-3D projection module to explicitly project the 2D features to the 3D point cloud region by region (centered at each correspondence pixel), according to the camera intrinsic and transformation matrix. We exhaustively explore the possible designs, e.g., implicitly concatenating per-pixel features to each point. Finally, we demonstrate that the design of explicitly projecting the deep color features in overlap-aware regions surpass implicit manner in our ablation studies.

Following Predator [23], we evaluate our work on 3DMatch and the more competitive and difficult 3DLoMatch [23] benchmark. In both benchmarks, we observe significant improvements in our proposed color and geometry learning strategy. Our approach outperforms the state-of-the-art method by a large margin of registration recall on the 3DLoMatch benchmark.

In summary, the contributions of our work are three-fold:

- We introduce a novel 2D-3D projection module that explicitly embeds the 2D color into the point cloud for registration task.
- We experimentally show that our method outperforms the baseline by a significant gap of 6.5% registration recall with 5000 sampled points on the more challenging 3DLoMatch benchmark.
- We conduct empirical studies and show the transfer ability of 2D pre-trained weights for 3D point cloud registration tasks.

## 2 Related Work

Advancements in deep learning enable fast development in many high-level and low-level tasks. In this section, we firstly review point cloud and image registration tasks, and then discuss a few additional relevant works in the area of multi-modal learning across color and geometry.

**Point Cloud Registration** Point Cloud Registration plays an important role in the computer vision community. Most successful methods in this field start with a low-level task, namely correspondence matching. A transformation matrix can

then be estimated from the predicted correspondences. Correspondences matching have been investigated even before deep learning era. Traditional machine learning methods and hand-crafted descriptors, such as ICP [4,8] and SIFT [28], have drawn great attention back then. Color ICP [30] leverages both color and geometry to align two partial point clouds, but with traditional machine learning optimizations. And the field is moving even faster since deep learning era. Leveraging the powerful deep learning features to learn rotation-invariant descriptors [10,6] for correspondences that are further fed into RANSAC for registration is the most successful story nowadays. Following the same pipeline, Predator [23] achieves the state-of-the-art results and first proposes to solve the registration problem on low-overlap frames. CoFiNet [46] proposes a coarse-to-fine manner on the point cloud to speed up the inference. GeoTransformer [34] leverages transformer on geometry to boost the point cloud registration. However, these prior work only use geometry as the single-source input. In this paper, we build upon their framework and propose an effective module that explicitly fuses the overlap regions learned from 2D color signals. BYOC [15] transfers the visual signals to train a geometric encoder. UnsupervisedR&R [14] uses differentiable rendering to enforce photometric and geometric consistency. Previous methods focus on self- or unsupervised learning on color signals. Our approach on the other hand discusses effectively making full use of RGB-D data as inputs.

**Image Registration** As the counterpart of 3D point cloud registration, 2D image registration contributes significantly to the computer vision community. It enables many high-level applications, such as 3D Reconstruction and visual SLAM [39,40]. Compared to point cloud registration, image registration uses only color input [38,35] and takes advantage of many existing pre-trained 2D network, such as ResNet with ImageNet pre-trained weights. The success of 2D image registration shows the possibility of learning registration from pixel input. Besides, the motivation of taking advantage of massive existing 2D pre-trained models suggests incorporating 2D signals into 3D registration task. In this work, we explore how to effectively use 2D signals on 3D registration task.

**2D-3D Multi-Modal Learning** Joint learning from color and geometry signals has been researched in many high-level tasks, such as in both 2D and 3D scene understanding [19,12,31,21,26,22,27]. 3D-SIS [19] proposes to implicitly leverage the color signal for 3D instance segmentation and detection tasks. RevalNet [20] adopts the similar idea of implicitly fusing color and geometry for 3D instance completion task. ImVoteNet [31] adds a 2D detector in addition to VoteNet [32] to explicitly use 2D color input. 3D-to-2D Distillation [27] presents a method to fuse 3D features for 2D semantic segmentation tasks. BPNet [22] uses a bidirectional projection module to mutually learn 2D-3D signals for both 2D and 3D semantic segmentation tasks. Besides scene understanding tasks, 2D-3D learning is also explored in representation learning. Pri3D [21] proposes to learn 2D representation in a pre-training paradigm for 2D scene understanding. P4Contrast [26] learns 3D representation from a novel 2D-3D loss for 3D scene understanding. Image2Point [45] boosts 3D point cloud understanding with 2D image pre-trained models. However, most of the previous research focus on high-

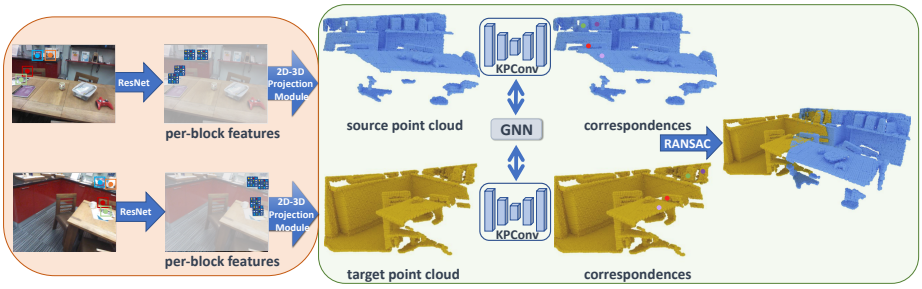


Fig. 2: **PCR-CG Pipeline.** The pipeline is composed of a 3D network, a 2D network and a 2D-3D projection module. Both 3D geometry and 2D images are taken as input and used to jointly learn features for detecting correspondences. The 2D network takes RGB images as input and extracts per-region features. A 2D-3D Projection Module is used to lift 2D pixel features into 3D point cloud. The concatenated features are fed into 3D network for finding correspondences. Due to our 2D-3D projection module, the 3D supervision can pass gradients back to the 2D network, and, therefore, yield an end-to-end training.

level semantic tasks. In this paper, we discuss the color-geometry learning in the point cloud registration task focusing on the low-level domain, i.e., predicting correspondences for point cloud registration. Additionally, we study the transfer ability of 2D pre-trained networks in the 3D registration task.

## 3 Methods

### 3.1 Data Representation

In our method, we use point cloud to represent the geometry input. At the same time, RGB images are the input to a pre-trained 2D network.

**Geometry Data** Each training sample contains a pair of non-aligned RGB-D frames. The transformation matrix that aligns them is used as the ground truth. We use point cloud lifted by depth frame as geometric input and predict the correspondences between them. For pre-computing the ground truth of correspondences, we transform one point cloud to the other, according to the aforementioned transformation matrix. Then, correspondences are found by a nearest neighbor search within a threshold in the Euclidean space.

**Color Data** We evaluate our method in RGB-D datasets, namely 3DMatch and 3DLoMatch. In these datasets, each point cloud is fused by 50 consecutive depth frames. The RGB images and depth images are in pairs. Therefore, each point cloud is also associated with 50 RGB frames. We pick up the first and the last RGB images for training and validation. Each RGB image is resized to the resolution of 240x320 in pixels. Notably, we do not need ground truth for 2D data, as the 2D network is pre-trained on other data.

### 3.2 Projection Module

**Insertion of the Projection Module** Before entering into our method, we have to revisit Predator [23]. In Predator, a point cloud is input to a 3D neural network. In the encoder, attention modules are used to correlate features obtained from source and target frames. The correlated features are fed into a decoder. The final layer outputs a score for each point to indicate its likelihood on overlapped regions. Per-point features from the final layer are used for finding correspondences. Next, correspondences are ranked based on the scores, and top-k correspondences' features are fed into RANSAC. Finally, RANSAC consumes the features of selected correspondences to further estimate the transformation matrix between the source and target frames. In this context, our module is directly inserted at the beginning of the 3D network without interfering the rest of the pipeline. The overview of the pipeline is illustrated in Fig. 2.

**Lifting 2D to 3D** To train on both color and geometry inputs, we propose a novel module that embeds deep color features into 3D representations. Our module PCR-CG takes the features extracted from RGB images and lifts them to 3D. The 3D network consumes a pair of point clouds, while our 2D-3D projection module takes the corresponding pairs of RGB images. To concatenate the features of 2D pixels into 3D points, we project XYZ coordinates of each point cloud onto its associated image planes. In our setup, we select the first and the last RBG images among 50 consecutive RGB-D frames that are used to generate the point cloud. Since each point cloud is tied to two color views, we average the feature vectors sampled from the overlapped regions. In the end, we append the feature vectors from 2D pixels to 3D points. We illustrate this projection procedure in Fig. 3. The 3D network remains the same as Predator [23] and we adjust input dimensions of the first layer. The combined features are fed into the KPConv encoder and are crossed at the bottleneck part via attention modules, which is identical to Predator [23].

**2D Pre-trained Networks** We empirically find appending RGB values to 3D points brings less gain. Similar results are observed in ImVoteNet [31] and 3DMV [12], and we also confirm this in the low-level task. Therefore, we propose to lift deep color features rather than RGB values. In our module, the 2D network is a standard ResUNet-50 backbone. We choose ResUNet since its encoder weights can be initialized by most popular 2D pre-trained models, such as ImageNet, Pri3D [21] and SuperGlue [37]. In this manner, we can easily change to different pre-trained networks. In ablation studies, we indicate that different 2D pre-trained weights have a significant influence on the 3D results.

**Frame Selection** Each point cloud is fused by 50 consecutive depth frames. We propose to use the first and last frames considering the performance and efficiency. Regarding the number of selected frames, we present the color coverage in Fig. 4. We show an increasing registration recall with more views in the ablation study.

**Implicit vs. Explicit Projection** Implicit projection lifts the features of every pixel, while the other projects features of some certain pixels in an explicit manner. In our design, we use a pre-trained 2D network, i.e., SuperGlue [37], to

predict the correspondences between the source and target RGB frames. Then, we project features extracted from the regions around the correspondences. In this manner, the regions are lifted explicitly to 3D, which indicates a rough overlap estimated from color signals. We experimentally demonstrate the advantages of explicit projection in the ablation study.

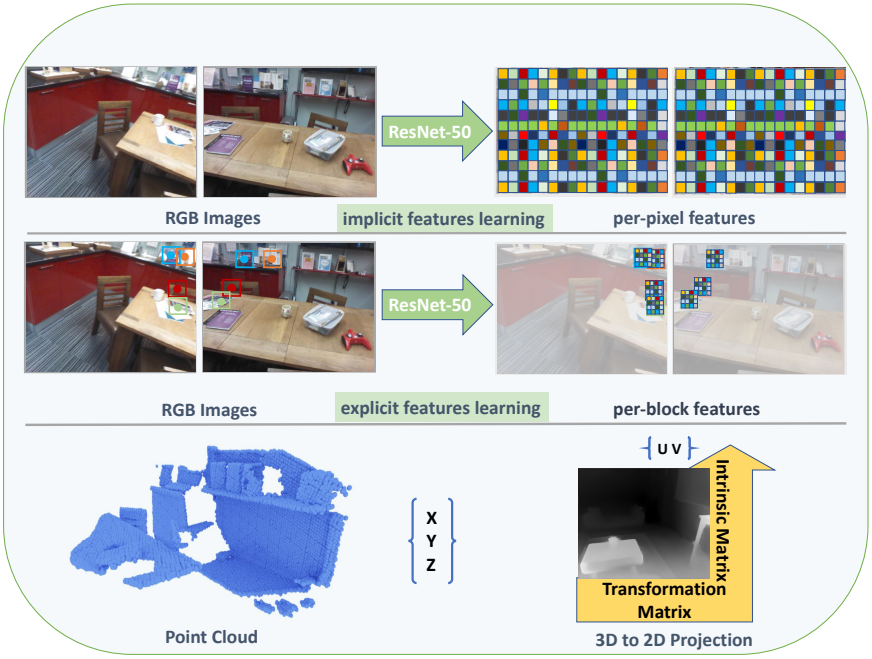


Fig. 3: PCR-CG: 2D-3D Projection Module. We introduce a novel 2D-3D Projection Module to lift 2D color features into 3D. The module takes the transformation matrix and depth map to project regional features to 3D point cloud.

## 4 Experimental Results

In this section, we show experimental results and ablation studies regarding the proposed 2D-3D projection module. We focus on indoor data, namely *3DMatch* and *3DLoMatch*. In the main result, different numbers of points are sampled for registration in RANSAC. Additionally, we conduct ablation studies, such as different projections and 2D pre-trained weights in Sec. 4.1.

**Experiments Setup** For training, we use the SGD optimizer with learning rate 0.005 and a batch size of 1. We use the exponential learning scheduler, and the learning rate is decreased by a factor of 0.95 in every epoch. During the

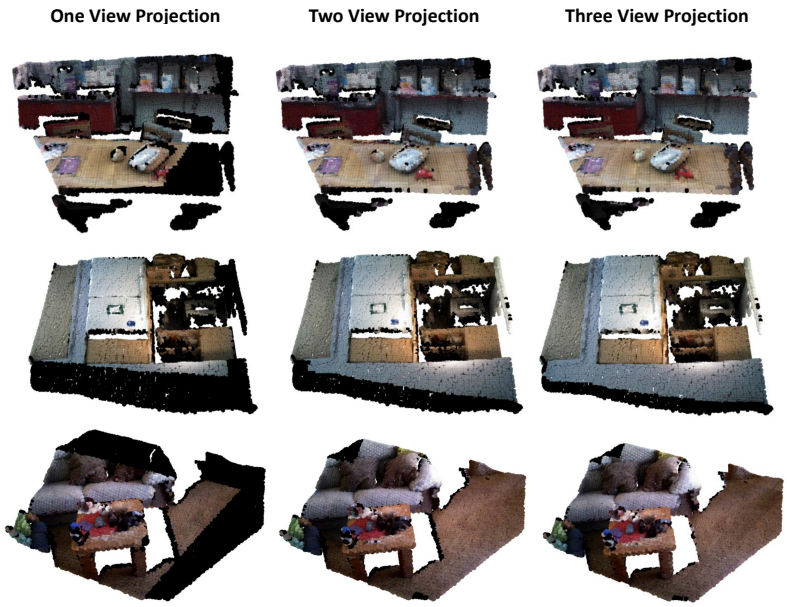


Fig. 4: Non-black points take 2D features; black points indicate no features taken from 2D. We observe not every point can be associated with the color features with one image; with two views, most points have the coverage of projected color features. However, adding the third view does not significantly improve cover coverage. Therefore, we choose two views as a default setup.

test, we use open3D for feature matching and RANSAC. For 2D networks, we use ResUNet-50 to extract per-pixel features. KPConv [44] is used as the 3D Backbone.

**Metrics** We mainly compare the results on four metrics, namely Registration Recall (RR), Feature Matching Recall (FMR), Relative Rotation Error (RRE), and Relative Translation Error (RTE). RR is the main metric we compare on and is most reliable, representing the fraction of pairs of point cloud, for which the correct transformation parameters are found after correspondence matching and RANSAC. Similar to Predator, we also report FMR, defined as the fraction of pairs that has at least 5% inlier matches. RTE and RRE measure the deviations from the ground truth pose. More specifically, RTE is computed by the differences between two frames by L1 norm; RRE is the drifted degrees between two frames registered by predicted transformation. Please refer to supplementary materials for detailed explanations and mathematical definitions.

**3DLoMatch** We show results on 3DLoMatch in Tab. 1. In 3DLoMatch, each pair of frames has at most 30% overlaps, and therefore it is a more challenging benchmark. We compare our method with SOTA methods in terms of RR and FMR. We show our method outperforms previous algorithms, including the most

| # Sampled Points | 3DMatch 5000 | 2500 | 1000 | 500 | 250 | 3DLoMatch 5000 | 2500 | 1000 | 500 | 250 |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Feature Matching Recall*(%) ↑ | | | | | | | | | |
| 3DSN [16] | 95.0 | 94.3 | 92.9 | 90.1 | 82.9 | 63.6 | 61.7 | 53.6 | 45.2 | 34.2 |
| FCGF [10] | 97.4 | 97.3 | 97.0 | 96.7 | 96.6 | 76.6 | 75.4 | 74.2 | 71.7 | 67.3 |
| D3Feat [6] | 95.6 | 95.4 | 94.5 | 94.1 | 93.1 | 67.3 | 66.7 | 67.0 | 66.7 | 66.5 |
| SpinNet [1] | 97.4 | 97.0 | 96.4 | 96.7 | 94.8 | 75.5 | 75.1 | 74.2 | 69.0 | 62.7 |
| Predator [23] | 96.6 | 96.6 | 96.5 | 96.3 | 96.5 | 78.6 | 77.4 | 76.3 | 75.7 | 75.3 |
| CoFiNet [46] | **98.1** | **98.3** | **98.1** | **98.2** | **98.3** | **83.1** | **83.5** | **83.3** | 83.1 | 82.6 |
| Ours – PCR-CG | 97.4 | 97.5 | 97.7 | 97.3 | 97.6 | 80.4 | 82.2 | 82.6 | **83.2** | **82.8** |
| | *Registration Recall*(%) ↑ | | | | | | | | | |
| 3DSN [16] | 78.4 | 76.2 | 71.4 | 67.6 | 50.8 | 33.0 | 29.0 | 23.3 | 17.0 | 11.0 |
| FCGF [10] | 85.1 | 84.7 | 83.3 | 81.6 | 71.4 | 40.1 | 41.7 | 38.2 | 35.4 | 26.8 |
| D3Feat [6] | 81.6 | 84.5 | 83.4 | 82.4 | 77.9 | 37.2 | 42.7 | 46.9 | 43.8 | 39.1 |
| SpinNet [1] | 88.8 | 88.0 | 84.5 | 79.0 | 69.2 | 58.2 | 56.7 | 49.8 | 41.0 | 26.7 |
| Predator [23] | 89.0 | 89.9 | **90.6** | 88.5 | 86.6 | 59.8 | 61.2 | 62.4 | 60.8 | 58.1 |
| CoFiNet [46] | 89.3 | 88.9 | 88.4 | 87.4 | **87.0** | **67.5** | 66.2 | 64.2 | 63.1 | 61.0 |
| Ours – PCR-CG | **89.4** | **90.7** | 90.0 | **88.7** | 86.8 | 66.3 | **67.2** | **69.0** | **68.5** | **65.0** |

Table 1: Results on *3DMatch* and *3DLoMatch*. Our holistic approach combining explicit deep color and geometric features results in significantly improved results over previous approaches including the most recent CoFiNet. PCR-CG surpasses our baseline Predator [23] by a large margin. Note that our approach uses the same backbone and pipeline as Predator and does not include the coarse-to-fine technique compared to CoFiNet [46]. Pri3D pre-trained model and two-view projection (explicit) are used for our approach.

recent CoFiNet [46] on different numbers of sampled points. More specifically, PCR-CG surpasses our baseline Predator by a large margin, especially with less sampled points, e.g., +7.7% on RR and +7.5% on FMR respectively with 500 sampled points. In Tab. 2, our approach outperforms previous methods also on Relative Rotation and Translation Errors. Besides the quantitative results, we show qualitative results in 3DLoMatch benchmark in Fig. 6.

**3DMatch.** We additionally report numbes on 3DMatch benchmark, where the overlap between two frames is at least 30%. Compared to 3DLoMatch, 3DMatch is easier and saturated. In Tab. 1, our proposed method outperforms our baseline Predator in both RR and FMR in most cases. Our method surpasses all the other methods, including the most recent CoFiNet, except for 5000 sampled points in FMR and 1000 sampled points in RR, where we achieve the second best number. In Tab. 2, our method also achieves SOTA results on Relative Rotation Error and Relative Translation Error.

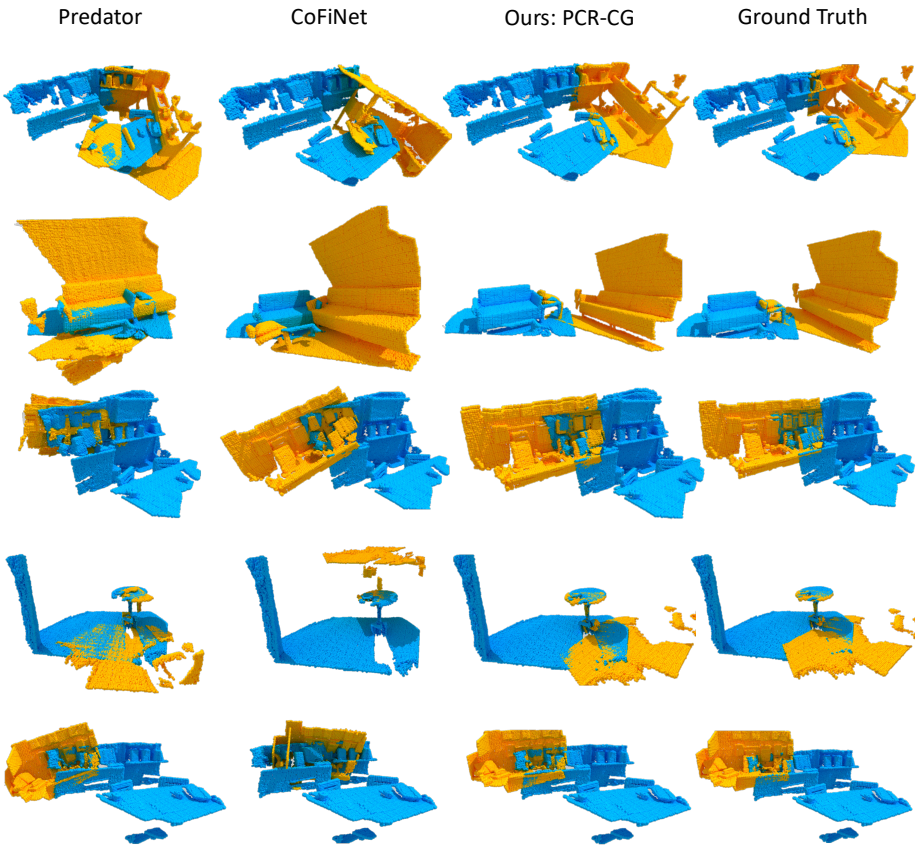


Fig. 5: Qualitative Comparisons on *3DLoMatch*. With the help of our proposed 2D-3D projection module, PCR-CG outperforms SOTA methods, such as our baseline Predator.

### 4.1 Ablation Study

In this section, we ablate the different designs of projecting 2D to 3D for point cloud registration task. We prove that our design of explicitly leveraging the color signals achieves the best result. Furthermore, we show the significant influence of the 2D pre-trained network and the frame selection on the final 3D registration results. We conduct our ablation experiments in *3DLoMatch* benchmark.

**Frame Selection and Color Coverage** In 3DLoMatch benchmark, each point cloud is fused by 50 consecutive frames. To ensure 100% color coverage, 50 frames must be all used for each point cloud. However, 50 times forward and backward passes are time-consuming. To ensure the color coverage as well as efficiency, we propose to use the first and last frame to back-project pixel features into 3D geometry. The visuals in Fig. 4 demonstrate that there are approximately 30%

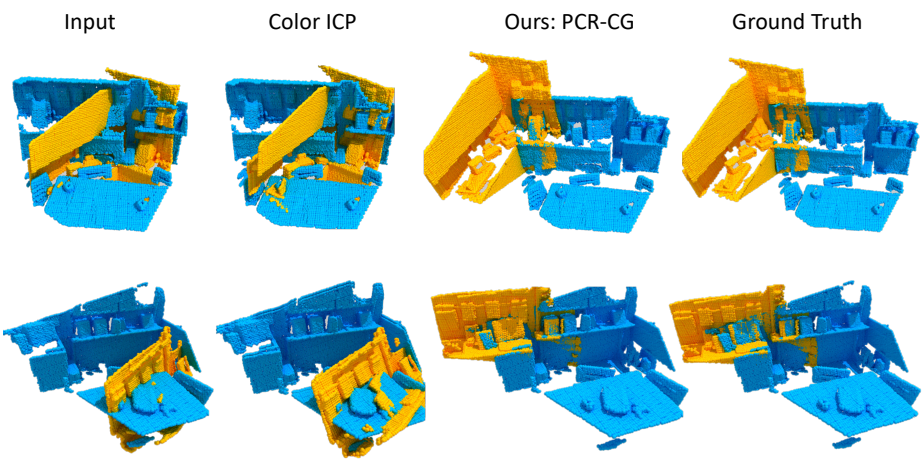


Fig. 6: Qualitative Comparisons on *3DLoMatch*. Compared to ColorICP [30], our method is more robust to initialization, especially in the case of large transformation and low overlaps such as in 3DLoMatch.

points that are not covered with one frame. With two views, most points are covered. With three views, it only slightly improves the color coverage. Quantitatively, the registration recall confirms the observation. To provide in-depth analysis of how much influence it has in terms of the number of views, we adjust the numbers of images used in the training and test. In Tab. 7, we can clearly see the signal that using more views leads to an increasing registration recall, which proves that the proposed 2D-3D Projection Module contributes significantly to the 3D registration task.

**Deep Color Features vs. RGBs** We compare our method to ColorICP [30] in Tab. 3 to show the effectiveness of deep color features. Similar to ICP, ColorICP also requires a good pose initialization. With original pose initialization that has a large transformation, both ICP and ColorICP failed to align two point clouds. With improved poses estimated by our method as initialization, both show marginal improvements. This observation demonstrates the importance of our method by embedding deep rather than shallow color features into geometry. Similarly, we use SIFT estimated from RGB images to register their point clouds. This result indicates the same conclusion and is showed in Tab. 5.

**2D Pre-trained Networks** PCR-CG lifts the 2D features into 3D. Therefore, massive existing 2D pre-trained models can be used. In Tab. 4, we show the influence of 2D representation on 3D results. We ablate on 2D models, such as ImageNet [13] and Pri3D [21] pre-trained weights. With 2D pre-trained weights, we achieve better 3D results compared to random initialization (Scratch). In general, we notice the trend that our method can achieve better registration recall numbers with more powerful 2D pre-trained models.

| | *3DMatch* | | *3DLoMatch* | |
|---|---|---|---|---|
| | RRE (°) | RTE (m) | RRE (°) | RTE (m) |
| 3DSN [16] | 2.199 | 0.071 | 3.528 | 0.103 |
| FCGF [10] | **1.949** | 0.066 | 3.146 | 0.100 |
| D3Feat [6] | 2.161 | 0.067 | 3.361 | 0.103 |
| Predator [23] | 2.029 | <u>0.064</u> | <u>3.048</u> | 0.093 |
| CoFiNet [46] | 2.002 | <u>0.064</u> | 3.271 | <u>0.090</u> |
| Ours – PCR-CG | <u>1.993</u> | **0.061** | **3.002** | **0.087** |

Table 2: Relative Rotation Errors (RRE) and Relative Translation Errors (RTE) with 5,000 sampled points on 3DMatch and 3DLoMatch benchmarks. Our approach achieves the best in RTE, and the second best in RRE.

| | Original Initial Pose | | Improved Initial Pose | |
|---|---|---|---|---|
| | *3DMatch* | *3DLoMatch* | *3DMatch* | *3DLoMatch* |
| ICP [36] | 4.20 | 1.40 | 91.0 | 68.5 |
| ColorICP [30] | 4.90 | 1.50 | **91.4** | **68.8** |
| PCR-CG | **90.7** | **68.2** | – | – |

Table 3: Registration Recall on *3DMatch* and *3DLoMatch*. Our method outperforms ICP and ColorICP on both benchmarks by large margins and more robust to the bad pose initialization.

**Window size.** We ablate different window sizes for extracting deep color features, and empirically find window size 11x11 achieves the best performance (see Tab. 6).

**Implicit vs. Explicit Projection** Adding color to 3D is non-trival. As aforementioned, we explore different ways of projecting 2D into 3D. Implicit one projects all the pixel values/features onto 3D, while explicit one leverages the 2D overlap information to project features region by region. We experimentally show that our design outperforms the rest. In Tab. 8, we show different combinations of 2D pre-trained weights and projections. In general, projecting deep color features such as Pri3D outperforms SIFT features and RGB values. In addition, we show that explicit projection outperforms implicit projection.

**Different Baselines** We adopt the same backbone and pipeline as Predator. However, our module is not specifically tied to Predator. Notably, our module is agnostic to methods, and it is easy to be plugged into any frameworks operating on RGB-D data. In Tab. 9, we demonstrate that our module also brings a significant improvement on CoFiNet baseline, i.e., +3.5% Registration Recall at 5,000 sampled points.

| | *2D Backbone* | *Registration Recall (%)* | | | | |
|---|---|---|---|---|---|---|
| | | 5000 | 2500 | 1000 | 500 | 250 |
| PCR-CG | Scratch | 66.1 | 67.2 | 68.2 | 68.3 | 64.7 |
| PCR-CG | ImageNet | 66.3 | **67.9** | 68.9 | 66.1 | 65.0 |
| PCR-CG | Pri3D | **66.3** | 67.2 | **69.0** | **68.5** | **65.0** |

Table 4: Ablation study on different 2D pre-trained models. We observe a clear correlation between 2D pre-trained weights and 3D results when explicitly lifting deep color features. Using 2D pre-trained weights indicates higher registration recalls. Two-view project is used. Note that all 2D models are pre-trained on other data, thus showing a strong transfer ability in our method.

| | 3DLoMatch | 3DMatch |
|---|---|---|
| SIFT-DLT | 0.4 | 0.9 |

Table 5: We utilize OpenCV SIFT-DLT [28] to calculate relative image pose and corresponding registration recall on 3DLoMatch benchmark.

| | *Window size* | *Registration Recall (%)* | | | | |
|---|---|---|---|---|---|---|
| | | 5000 | 2500 | 1000 | 500 | 250 |
| PCR-CG | 3x3 | 63.1 | 64.0 | 65.1 | 64.6 | 60.7 |
| PCR-CG | 7x7 | 64.7 | 66.4 | 67.1 | 65.8 | 62.6 |
| PCR-CG | 11x11 | **66.3** | **67.2** | **69.0** | **68.5** | **65.0** |
| PCR-CG | 17x17 | 64.1 | 65.7 | 66.2 | 65.6 | 62.1 |

Table 6: Ablation study on window sizes in 3DLoMatch. We empirically found $11x11$ window size shows the best performance. Two-view projection is used. SuperGlue [37] is used to find correspondences and Pri3D [21] pre-trained model is used for feature extraction.

| | *Views* | *Registration Recall (%)* | | | | |
|---|---|---|---|---|---|---|
| | | 5000 | 2500 | 1000 | 500 | 250 |
| PCR-CG | 1 | 64.4 | 67.0 | 66.6 | 66.4 | 64.3 |
| PCR-CG | 2 | 66.3 | 67.2 | 69.0 | 68.5 | 65.0 |
| PCR-CG | 3 | **66.7** | **67.9** | **69.1** | **68.7** | **65.1** |

Table 7: Ablation study on color coverage. We show an increasing registration recall with more views used. SuperGlue [37] is used to find correspondences and Pri3D [21] pre-trained model is used for feature extraction.

| *Method* | *Features* | *Projection* | *Registration Recall (%)* | | | | |
|---|---|---|---|---|---|---|---|
| | | | 5000 | 2500 | 1000 | 500 | 250 |
| PCR-CG | RGB | implicit | 60.5 | 63.0 | 63.6 | 62.3 | 59.4 |
| PCR-CG | RGB | explicit | 60.4 | 63.1 | 63.5 | 62.8 | 59.9 |
| PCR-CG | SIFT | implicit | 63.1 | 65.1 | 65.5 | 64.9 | 61.4 |
| PCR-CG | SIFT | explicit | 64.8 | 67.0 | 67.1 | 66.5 | 63.9 |
| PCR-CG | SuperGlue | explicit | 64.0 | 65.0 | 65.0 | 65.0 | 60.8 |
| PCR-CG | ImageNet | implicit | 63.2 | 65.4 | 65.7 | 64.9 | 61.1 |
| PCR-CG | ImageNet | explicit | 66.3 | **67.9** | 68.9 | 66.1 | 65.0 |
| PCR-CG | Pri3D | implicit | 63.4 | 65.4 | 66.0 | 65.2 | 61.4 |
| PCR-CG | Pri3D | explicit | **66.3** | 67.2 | **69.0** | **68.5** | **65.0** |

Table 8: Ablation study on projections. RGB means simply appending RGB colors to point cloud. SIFT refers to projecting SIFT features onto points. Pri3D uses pre-trained weights to extract per-pixel features and projects them onto points. Similarly, SuperGlue/ImageNet refers to projecting SuperGlue/ImageNet pre-trained features. We show projecting deep color features outperforms SIFT and RGB values with the same projection. Implicit manner projects features of every pixel onto 3D, while explicit one projects features based on correspondences estimated by SuperGlue. We demonstrate the explicit projection surpasses the implicit one. Two-view projection is used in the experiments.

| *Baseline Method* | *Registration Recall (%)* | | | | |
|---|---|---|---|---|---|
| | 5000 | 2500 | 1000 | 500 | 250 |
| CoFiNet | 67.5 | 66.2 | 64.2 | 63.1 | 61.0 |
| CoFiNet (re-train) | 64.4 | 64.2 | 63.1 | 62.1 | 59.8 |
| CoFiNet + PCR-CG | **67.9** | **67.0** | **65.4** | **64.2** | **62.2** |

Table 9: Registration Recall based on CoFiNet on 3DLoMatch benchmark. We can notice a clear gap of plugging in our module compared to CoFiNet baseline. In this ablation experiment, our implementation is built upon the officially released code of CoFiNet. Thus, we re-train the official released code for a fair comparison. Pri3D pre-trained model and two-view projection (explicit) are used.

## 5 Conclusion

In this work, we correlate color and geometry for point cloud registration. To fully leverage RGB-D data, we propose a novel 2D-3D projection module to explicitly lift 2D features into 3D. Our module enables the usage of massive existing 2D pre-trained networks in 3D registration tasks. We hope our research can inspire the community to pay more attention on joint learning with color and geometry on various computer vision applications.

**Acknowledgments** This work is supported by the Joint Funds of Zhejiang NSFC (LTY22F020001) and Open Research Fund of State Key Laboratory of Transient Optics and Photonics. Yu Zhang is the corresponding author.

## References


1. Ao, S., Hu, Q., Yang, B., Markham, A., Guo, Y.: Spinnet: Learning a general surface descriptor for 3d point cloud registration. In: CVPR. pp. 11753–11762 (2021) 9
2. Aoki, Y., Goforth, H., Srivatsan, R.A., Lucey, S.: Pointnetlk: Robust & efficient point cloud registration using pointnet. In: CVPR. pp. 7163–7172 (2019) 2
3. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3D semantic parsing of large-scale indoor spaces. In: ICCV (2016) 2
4. Arun, K.S., Huang, T.S., Blostein, S.D.: Least-squares fitting of two 3-d point sets. TPAMI (5), 698–700 (1987) 4
5. Bai, X., Luo, Z., Zhou, L., Chen, H., Li, L., Hu, Z., Fu, H., Tai, C.L.: Pointdsc: Robust point cloud registration using deep spatial consistency. In: CVPR. pp. 15859–15869 (2021) 2
6. Bai, X., Luo, Z., Zhou, L., Fu, H., Quan, L., Tai, C.L.: D3feat: Joint learning of dense detection and description of 3d local features. In: CVPR. pp. 6359–6367 (2020) 4, 9, 12
7. Balntas, V., Doumanoglou, A., Sahin, C., Sock, J., Kouskouridas, R., Kim, T.K.: Pose guided rgbd feature learning for 3d object pose estimation. In: CVPR. pp. 3856–3864 (2017) 2
8. Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. In: Sensor fusion IV: control paradigms and data structures. vol. 1611, pp. 586–606. International Society for Optics and Photonics (1992) 4
9. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. arXiv preprint arXiv:1709.06158 (2017) 2
10. Choy, C., Park, J., Koltun, V.: Fully convolutional geometric features. In: CVPR. pp. 8958–8966 (2019) 4, 9, 12
11. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3D reconstructions of indoor scenes. In: CVPR (2017) 2
12. Dai, A., Nießner, M.: 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In: ECCV. pp. 452–468 (2018) 2, 4, 6
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 11
14. El Banani, M., Gao, L., Johnson, J.: Unsupervisedr&r: Unsupervised point cloud registration via differentiable rendering. In: CVPR. pp. 7129–7139 (2021) 2, 4
15. El Banani, M., Johnson, J.: Bootstrap your own correspondences. In: ICCV. pp. 6433–6442 (2021) 4
16. Gojcic, Z., Zhou, C., Wegner, J.D., Wieser, A.: The perfect match: 3d point cloud matching with smoothed densities. In: CVPR. pp. 5545–5554 (2019) 9, 12
17. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017) 2
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 2
19. Hou, J., Dai, A., Nießner, M.: 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In: CVPR (2019) 4
20. Hou, J., Dai, A., Nießner, M.: RevealNet: Seeing Behind Objects in RGB-D Scans. In: CVPR (2020) 4
21. Hou, J., Xie, S., Graham, B., Dai, A., Nießner, M.: Pri3d: Can 3d priors help 2d representation learning? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5693–5702 (2021) 2, 3, 4, 6, 11, 13

22. Hu, W., Zhao, H., Jiang, L., Jia, J., Wong, T.T.: Bidirectional projection network for cross dimension scene understanding. In: CVPR. pp. 14373–14382 (2021) 4
23. Huang, S., Gojcic, Z., Usvyatsov, M., Wieser, A., Schindler, K.: Predator: Registration of 3d point clouds with low overlap. In: CVPR. pp. 4267–4276 (June 2021) 2, 3, 4, 6, 9, 12
24. Lahoud, J., Ghanem, B., Pollefeys, M., Oswald, M.R.: 3d instance segmentation via multi-task metric learning. In: ICCV (2019) 2
25. Liu, Y., Fan, Q., Zhang, S., Dong, H., Funkhouser, T., Yi, L.: Contrastive multimodal fusion with tupleinfonce. In: CVPR. pp. 754–763 (2021) 2
26. Liu, Y., Yi, L., Zhang, S., Fan, Q., Funkhouser, T., Dong, H.: P4contrast: Contrastive learning with pairs of point-pixel pairs for rgb-d scene understanding. arXiv preprint arXiv:2012.13089 (2020) 2, 4
27. Liu, Z., Qi, X., Fu, C.W.: 3d-to-2d distillation for indoor scene parsing. In: CVPR. pp. 4464–4474 (2021) 4
28. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2), 91–110 (2004) 4, 13
29. Niethammer, M., Kwitt, R., Vialard, F.X.: Metric learning for image registration. In: ICCV. pp. 8463–8472 (2019) 2
30. Park, J., Zhou, Q.Y., Koltun, V.: Colored point cloud registration revisited. In: ICCV. pp. 143–152 (2017) 4, 11, 12
31. Qi, C.R., Chen, X., Litany, O., Guibas, L.J.: Imvotenet: Boosting 3D object detection in point clouds with image votes. In: CVPR (2020) 2, 4, 6
32. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: ICCV. pp. 9277–9286 (2019) 4
33. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3D classification and segmentation. CVPR (2017) 2
34. Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Xu, K.: Geometric transformer for fast and robust point cloud registration. In: CVPR. pp. 11143–11152 (2022) 4
35. Revaud, J., Weinzaepfel, P., De Souza, C., Pion, N., Csurka, G., Cabon, Y., Humenberger, M.: R2d2: repeatable and reliable detector and descriptor. arXiv preprint arXiv:1906.06195 (2019) 4
36. Rusinkiewicz, S., Levoy, M.: Efficient variants of the icp algorithm. In: Proceedings third international conference on 3-D digital imaging and modeling. pp. 145–152. IEEE (2001) 12
37. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: CVPR (2020) 2, 3, 6, 13
38. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: CVPR. pp. 4938–4947 (2020) 4
39. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016) 4
40. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: ECCV (2016) 4
41. Song, S., Xiao, J.: Sliding shapes for 3D object detection in depth images. In: ECCV (2014) 2
42. Srinivasan, P.P., Wang, T., Sreelal, A., Ramamoorthi, R., Ng, R.: Learning to synthesize a 4d rgbd light field from a single image. In: CVPR. pp. 2243–2251 (2017) 2
43. Stückler, J., Gutt, A., Behnke, S.: Combining the strengths of sparse interest point and dense image registration for rgb-d odometry. In: ISR/Robotik; International Symposium on Robotics. pp. 1–6. VDE (2014) 2

44. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: KPConv: Flexible and deformable convolution for point clouds. In: CVPR (2019) 8
45. Xu, C., Yang, S., Zhai, B., Wu, B., Yue, X., Zhan, W., Vajda, P., Keutzer, K., Tomizuka, M.: Image2point: 3d point-cloud understanding with 2d image pretrained models (2021) 4
46. Yu, H., Li, F., Saleh, M., Busam, B., Ilic, S.: Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. NeurIPS **34** (2021) 2, 4, 9, 12
47. Zeng, A., Song, S., Nießner, M., Fisher, M., Xiao, J., Funkhouser, T.: 3DMatch: Learning local geometric descriptors from RGB-D reconstructions. In: CVPR (2017) 2
48. Zhou, Q., Sattler, T., Leal-Taixe, L.: Patch2pix: Epipolar-guided pixel-level correspondences. In: ICCV. pp. 4669–4678 (2021) 2