SemReg: Semantics Constrained Point Cloud Registration

Sheldon Fung[1], Xuequan Lu*[1], Dasith de Silva Edirimuni[2], Wei Pan[3], Xiao Liu[2], and Hongdong Li[4]

[1] La Trobe University, Australia
[2] Deakin University, Australia
[3] OPTMV, China
[4] Australian National University, Australia

Abstract. Despite the recent success of Transformers in point cloud registration, the cross-attention mechanism, while enabling point-wise feature exchange between point clouds, suffers from redundant feature interactions among semantically unrelated regions. Additionally, recent methods rely only on 3D information to extract robust feature representations, while overlooking the rich semantic information in 2D images. In this paper, we propose SemReg, a novel 2D-3D cross-modal framework that exploits semantic information in 2D images to enhance the learning of rich and robust feature representations for point cloud registration. In particular, we design a Gaussian Mixture Semantic Prior that fuses 2D semantic features across RGB frames to reveal semantic correlations between regions across the point cloud pair. Subsequently, we propose the Semantics Guided Feature Interaction module that uses this prior to emphasize the feature interactions between the semantically similar regions while suppressing superfluous interactions during the cross-attention stage. In addition, we design a Semantics Aware Focal Loss that facilitates the learning of robust features, and a Semantics Constrained Matching module that performs matching only between the regions sharing similar semantics. We evaluate our proposed SemReg on the public indoor (3DMatch) and outdoor (KITTI) datasets, and experimental results show that it produces superior registration performance to state-of-the-art techniques. Code is available at: `https://github.com/SheldonFung98/SemReg.git`

Keywords: Point Cloud Registration · RGB-D · Cross-modality

1 Introduction

The task of estimating the relative transformation between two point clouds, often referred to as point cloud registration, is a fundamental problem in 3D computer vision. With the proliferation of consumer 3D imaging devices [53], captured point clouds have increasingly been used in numerous tasks, such as

* Corresponding author.

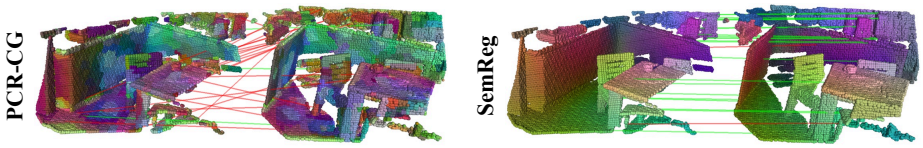

Fig. 1: Comparison between the point-wise features and correspondences by the cross-modal method PCR-CG [52] and our proposed SemReg. We visualize the features by applying T-SNE on the final point-wise features. Unlike PCR-CG, our method generates semantically consistent features as evidenced by color consistency. We visualize established correspondences with green lines for inliers and red lines for outliers.

denoising [35, 36, 51], segmentation [32], and detection [11, 33]. Since it is rarely possible to capture an entire scene in a single point cloud, the task of aligning multiple overlapping point clouds obtained from a given scene, known as registration, has become highly relevant [22, 23].

While early approaches for point cloud registration relied on handcrafted features [9, 30], recent state-of-the-art methods have shifted their focus towards learning consistent feature representations [17, 20, 27]. Typically, these methods employ convolution networks to extract local geometry-based features, followed by multiple attention layers to aggregate global context information. Recent efforts concentrate on incorporating point-wise relative spatial information [20] or rotationally invariant features [27]. However, in practice, semantically unrelated regions are less likely to share overlap. Redundant feature interactions between such regions would lead to erroneous matching. While extracting semantic features from 3D geometry remains a difficult problem, huge advances have been made in the 2D image domain [26]. PCR-CG [52] is a 2D-3D cross-modal method that embeds the features learned from the 2D domain into 3D geometry-based features. Despite some improvement in registration success, it suffers from subpar matching due to the lack of specific supervision for optimizing 2D feature extraction (see Figure 1).

We observe that reliable correspondences naturally exist between two semantically consistent regions for point cloud registration. For instance, considering a point cloud pair where the overlapping region includes a table and a chair, valid correspondences typically exist only between regions representing tables or between regions representing chairs. Conversely, searching correspondences between table and chair is less fruitful. Motivated by this, we propose *SemReg*, a novel 2D-3D cross-modal point cloud registration approach that leverages semantic information to learn robust feature representations. In particular, we design a Gaussian Mixture Semantic Prior (GMSP) that fuses 2D semantic features across RGB frames to eliminate unnecessary interactions between semantically unrelated regions in the cross-attention stage. To achieve this, given the source and target point clouds and their respective RGB images, we first extract and project the high-level semantic information from the 2D images onto 3D space. Then, for each point in the source point cloud, we compute the semantic similarity with all points in the target point cloud and select the salient semantic points.

With these salient points as centers, we compute Gaussian distributions based on the point-wise Euclidean distance in the target point cloud. The GMSP is finally calculated as the mixture of the resulting Gaussian distributions. The GMSP is illustrated in Figure 2. We then propose the Semantics Guided Feature Interaction (SGFI) module that uses GMSP to identify the underlying semantics of points within the target point cloud, given the initial query point in the source point cloud. Subsequently, this prior is incorporated into the cross-attention stage to eliminate redundant interactions among semantically unrelated regions. By doing so, our method yields more robust features, leading to better correspondences as compared to PCR-CG [52], a previous cross-modal method (see Figure 1). In addition, we design a Semantics Aware Focal Loss that leverages the cross-point-cloud semantic similarity to emphasize feature interactions within semantically similar regions while eliminating the remaining interactions within other regions. Also, we develop a Semantics Constrained Matching module to exclusively perform matching between the regions sharing similar semantics.

We evaluate our proposed SemReg on the indoor 3DMatch [50] and outdoor KITTI [13] datasets, and experimental results show that it outperforms state-of-the-art methods. Fine alignment results showcase the robustness of the learned features, achieving a final registration rate of 95.4% and 77.4% for 3DMatch and 3DLoMatch, respectively. Our contributions are summarized as follows:

- We propose SemReg, a novel 2D-3D cross-modal framework that exploits semantic information in images to enhance the learning of rich and robust feature representations for point cloud registration.
- We introduce a Semantics Guided Feature Interaction module incorporating the designed Gaussian Mixture Semantic Prior to eliminate the feature interactions between semantically unrelated regions.
- We design a Semantics Aware Focal Loss that further emphasizes feature interactions within semantically similar regions. We also develop Semantics Constrained Matching that constrains the feature matching to be conducted between regions sharing high semantic correlations.

2 Related Work

2.1 Point Cloud Registration

Correspondence-based methods. Estimating the alignment transformation between two point clouds from reliable correspondences is the most common approach to point cloud registration [1, 6, 10, 12, 14, 19, 24, 34, 39, 40, 45, 48]. In general, these methods follow a four-step pipeline: 1) extract point-wise features, 2) form correspondences from features, 3) remove outliers (typically using RANSAC [31]), and 4) compute the transformation with Singular Value Decomposition (SVD). Pioneering works extract features from the local geometry. Zeng *et al.* [50] employed a Siamese network to extract features from volumetric blocks randomly sampled from point cloud pairs. Building on this, Gojcic

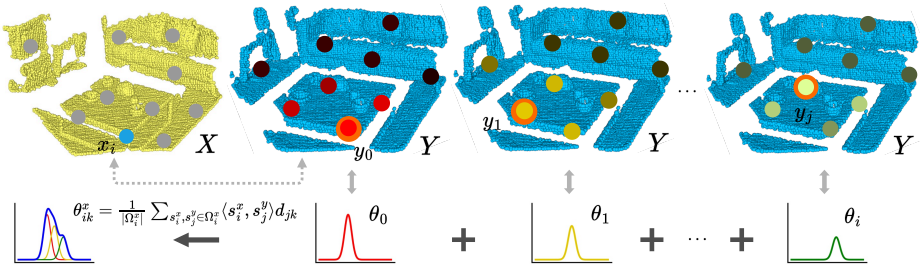

Fig. 2: Illustration of our proposed GMSP. Given a reference point (blue dot) in the point cloud $\mathbf{X}$, we select the semantically salient points in $\mathbf{Y}$ (indicated by the brightest points in $\mathbf{Y}$ with orange outer circle). With these points as centers, we compute the Gaussian distribution over the point cloud $\mathbf{Y}$ based on Euclidean distance between points. The GMSP (blue curve on the rightmost) is the mixture of these distributions.

et al. [15] proposed the Smoothed Density Value (SDV) and leveraged the Local Reference Frame (LRF) to improve generalization ability. Despite promising results, performance is still constrained by the use of randomly selected local feature descriptors. This limitation arises from the need to balance computational efficiency and reproducibility. In an attempt to overcome the issue, Deng *et al.* [7, 8] employed Point Pair Features (PPF) [9] to enhance feature discriminability. To process large scene data, Bai *et al.* [3] adopted KPConv [37] to extract local feature descriptors.

The attention mechanism [38] has proven to be significant in various fields including natural language processing [28] and 2D/3D visual data processing [25, 54]. Predator [17] first endeavored to incorporate self/cross attention layers to solve two existing problems: 1) limited perception fields of the convolution operation, and 2) the absence of cross-point-cloud interaction. Building on this work, Lepard [20], introduced by Li *et al.*, employed a rotary positional encoding to reveal the point-wise distance during the attention operation. Similarly, Qin *et al.* [27] proposed GeoTransformer, a geometric self-attention module to strengthen the model's robustness against rotational transformations.

Direct registration methods. Another line of research involves estimating the transformation in an end-to-end fashion [2, 18, 21, 42–44]. Building on the Iterative Closest Point (ICP) method [4], substantial research has been dedicated to employing deep learning techniques to predict soft correspondences iteratively [21, 42, 43, 46]. These approaches effectively mitigate the sub-optimal local minima problem within the ICP method. Aoki *et al.* [2] modified the Lucas & Kanade (LK) algorithm and adapted it to point cloud registration. Xu *et al.* [44] proposed a network architecture to simultaneously update both the predicted transformations and overlapping masks. Exploiting the inference power of the Transformer, Yew *et al.* [47] proposed to directly regress the final set of sparse correspondence locations, allowing transformation prediction without RANSAC.

2D-3D cross-modal methods. In addition to depth information, RGB-D scans also provide point-wise color information, which is usually neglected by most methods. PCR-CG [52] pioneered the successful integration of this color information by projecting and concatenating the 2D color channels with the point-wise features. In contrast, we explicitly extract high-level semantic features from corresponding 2D images and leverage them as prior information within an innovative cross-modal framework that is able to learn more robust 3D features.

3 Motivation

Correspondence matching between points across point clouds is crucial to accurate registration yet two key research gaps limit current methods: 1) improving feature similarity between salient, corresponding points across point clouds by incorporating associated point semantics and 2) fully exploiting cross-modal 2D-3D information. Our novel approach addresses both gaps by considering invariant 2D features from RGB frames to enrich the respective 3D features with rich semantic information. We further analyze these gaps below:

1) Improving feature similarity across points via semantics. To impose a high degree of feature similarity, the feature output from the network should ideally remain invariant. However, current methods solely focus on the geometric properties of points and disregard semantic information inherent to them when attempting to effectuate feature invariance. For example, Lepard [20] proposes a rotary positional encoding to uncover relative point position information which is then exploited within a transformer architecture to create position-consistent features. Nevertheless, this encoding solely relies on Euclidean distances between points. This may lead to erroneous matching due to ambiguous features as geometric structure alone omits crucial semantic information. For instance, points within partial scans of a bed and table may exhibit geometric similarities but should produce distinct features due to different semantics.

2) Fully exploiting cross-modal 2D-3D information. We observe that current cross-modal methods, such as the pioneering work of PCR-CG [52], only use color information within RGB images when aligning the associated point clouds via registration. Similar to 1) above, semantic information within the images is not leveraged which significantly curtails performance.

We intuit that both issues are interconnected, i.e., determining point semantics requires a high degree of feature invariance which is not directly available to 3D point convolution methods but may be accessible through 2D image convolution advances. Therefore, we pioneer a cross-modal approach to exploit rich, invariant 2D features of RGB frames produced by a pre-trained lightweight DinoV2 backbone [26]. These 2D features uncover high-level point semantics which reinforce the final 3D features and lead to better correspondences. See Sec. 4.3 for more details. As illustrated in Figure 1, our method produces more consistent point features that are semantically accurate, leading to better correspondences, while PCR-CG generates less consistent features and poor correspondences.

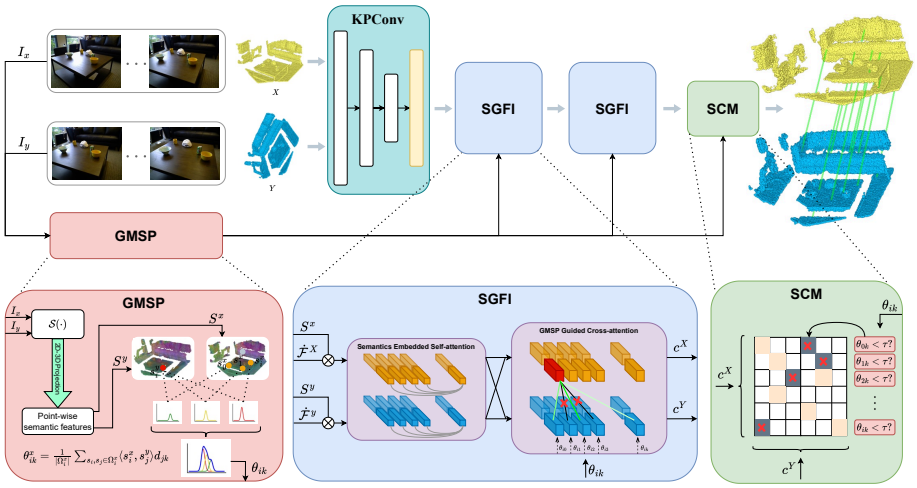

Fig. 3: Our proposed SemReg. KPConv is employed to simultaneously down-sample the point cloud into multiple resolution levels and extract point-wise local feature descriptors. The proposed Gaussian Mixture Semantic Prior module (GMSP) generates semantic priors θ from 2D images. Both semantic priors and local features are then fed into our Semantics Guided Feature Interaction module (SGFI). Subsequently, the Semantics Constrained Matching module (SCM) establishes the final correspondences.

4 Method

4.1 Problem Definition

Given two partially overlapping point clouds $\mathbf{X} \in \mathbb{R}^{m\times 3}$ and $\mathbf{Y} \in \mathbb{R}^{n\times 3}$ with their corresponding RGB images $I_x, I_y \in \mathbb{R}^{w\times h\times 3}$, our task is to establish a set of correspondences $\mathbb{C}^* = \{(\mathbb{C}_{x_i}, \mathbb{C}_{y_i}) \mid \mathbb{C}_{x_i} \in \mathbb{R}^3, \mathbb{C}_{y_i} \in \mathbb{R}^3, i = 1, \ldots, t\}$. Then the rigid transformation $\mathbf{T}_X^Y = \{\mathbf{R}^*, t^*\}$, with $\mathbf{R}^* \in SO(3)$ and $t^* \in \mathbb{R}^3$, which aligns $\mathbf{X}$ to $\mathbf{Y}$, can be obtained by solving the following equation:

$$\mathbf{R}^*, t^* = \min_{\mathbf{R},t} \sum_{(\mathbb{C}_{x_i}, \mathbb{C}_{y_i}) \in \mathbb{C}^*} \|\mathbf{R} \cdot \mathbb{C}_{x_i} + t - \mathbb{C}_{y_i}\|_2^2. \tag{1}$$

4.2 Local Feature Encoder

Typically, point clouds generated from RGB-D data are dense and noisy. We follow previous works [17, 20, 27, 47] and employ KPConv [37] to extract multi-level local-geometry features as follows:

$$\{\dot{\mathbf{X}}, \mathcal{F}^X\} = \kappa(X), \quad \{\dot{\mathbf{Y}}, \mathcal{F}^Y\} = \kappa(Y), \tag{2}$$

where $\dot{\mathbf{X}} \in \mathbb{R}^{m'\times 3}$, $\dot{\mathbf{Y}} \in \mathbb{R}^{n'\times 3}$, $\mathcal{F}^X \in \mathbb{R}^{m'\times d_k}$, and $\mathcal{F}^Y \in \mathbb{R}^{n'\times d_k}$ are the outputted down-sampled points and their corresponding features. m' and n' are

the numbers of the downsample points. The KPConv-like network $\kappa(\cdot)$ comprises a series of ResNet-based convolutional layers and multiple downsampling layers (detailed in Section 2.1 in the supplementary document). We designate the first layer in the decoder as the backbone output layer since the point number of this layer ($\sim 1.5K$) closely aligns with the number of semantic features that can be acquired for each point cloud ($\sim 2K$, as detailed in Section 4.3).

4.3 Semantics Guided Feature Interaction

Despite recent advances in neural architectures for 3D registration, extracting transformation invariant features remains a great challenge. Fortunately, recent work in 2D representation learning [26] has shown promise in generating invariant 2D feature representations. Motivated by this, we propose a Semantics Guided Feature Interaction (SGFI) module that hierarchically exploits the robust 2D features to determine underlying semantics that drive the 3D registration in a cross-modal manner. It consists of three components: 1) Gaussian Mixture Semantic Prior (GMSP), 2) Semantics Embedded Self-attention, and 3) GMSP Guided Cross-attention. The core idea of our method is to extract rich, invariant 2D features that carry semantic information and project them onto 3D space. These 2D-3D semantic features are first used as embeddings in the self-attention layer and further exploited to generate a Gaussian Mixture Semantic Prior (GMSP) that captures point-wise semantic information between source and target scenes. These priors are then used in the cross-attention step to enhance point features that are decoded to obtain correspondences.

Gaussian Mixture Semantic Prior. Given corresponding images $I_x, I_y \in \mathbb{R}^{w\times h\times 3}$, we use a semantic feature extractor $\mathcal{S}(\cdot)$ to obtain dense semantic features $S_x \in \mathbb{R}^{m'\times d_s}$ and $S_y \in \mathbb{R}^{n'\times d_s}$:

$$S^x = \mathcal{P}(\text{norm}(\mathcal{S}(I_x))),\ S^y = \mathcal{P}(\text{norm}(\mathcal{S}(I_y))), \tag{3}$$

where $\mathcal{P}(\cdot)$ is the 2D-to-3D projection function with the provided depth channel and camera parameters while $\text{norm}(\cdot)$ represents a $\mathcal{L}_2$ normalization operator. In our implementation, we adopt a pre-trained lightweight DINOv2 [26] backbone as the semantic feature extractor $\mathcal{S}(\cdot)$, which remains fixed during training. Note that the projected semantic features S^x, S^y are respectively aligned with the sparse point clouds $\dot{\mathbf{X}}$ and $\dot{\mathbf{Y}}$. As a result, the semantic features for $x_i \in \dot{\mathbf{X}}$ and $y_i \in \dot{\mathbf{Y}}$ are denoted as s_i^x and s_j^y, respectively. We design the Gaussian Mixture Semantic Prior to reveal the semantic correlation across two point clouds. Following common practices [17, 20, 27], point cloud pairs with an overlap ratio under 10% are excluded. We therefore assume that semantically related regions exist under such a setting. For point x_i, we compute the cosine similarity between s_i^x and each $s_j^y \in S^y$. We then select those point pairs with similarity over a threshold η to form a set of salient semantic correspondences Ω_i^x:

$$\Omega_i^x = \{(s_i^x, s_j^y) \mid \langle s_i^x, s_j^y \rangle > \eta, \forall s_j^y \in S^y\}. \tag{4}$$

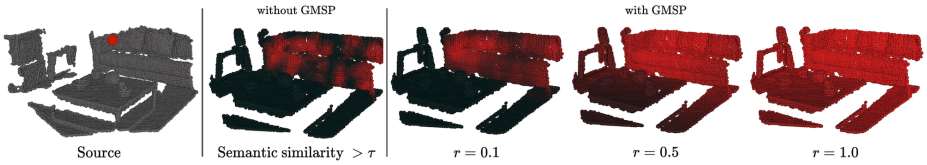

Fig. 4: The impact of our proposed GMSP. Given a point (red dot) in the source point cloud, the second figure from the left shows the result without the GMSP. Conversely, the three figures on the right side depict the semantic priors after integrating the Gaussian mixture with different support radius r. We can observe that semantics priors with the Gaussian mixture are smoother at the regions with high semantic similarity.

For each $s_j^y \in \Omega_i^x$, its corresponding y_j acts as the center of a Gaussian distribution. Then the Gaussian Mixture Semantic Prior $\theta_{ik}^x \in \mathbb{R}$ with respect to $x_i \in \dot{\mathbf{X}}$ and $y_k \in \dot{\mathbf{Y}}$ is defined as:

$$\theta_{ik}^x = \frac{1}{|\Omega_i^x|} \sum_{s_i^x, s_j^y \in \Omega_i^x} \langle s_i^x, s_j^y \rangle d_{jk}, \tag{5}$$

where d_{jk} is the Gaussian correlation matrix based on their Euclidean distance. Given a pair of points $y_j, y_k \in \dot{\mathbf{Y}}$, d_{jk} can be computed as $d_{jk} = \exp(-\|y_j - y_k\|_2^2 / r^2)$, where the hyper-parameter r is empirically set to 0.1. The impact of r is visualized in Figure 4. The overall computation process of the Gaussian Mixture Semantic Prior is visualized in Figure 2.

Semantics Embedded Self-attention. In the self-attention layer, query, key, and value triplets are obtained from the same point cloud. We explain the computation process using the point cloud $\mathbf{X}$ only. We first fuse the local feature $f_i^x \in \mathcal{F}^X$ with its corresponding 2D semantic feature:

$$f_i^x \leftarrow f_i^x + s_i^x W_s, \tag{6}$$

where $W_s \in \mathbb{R}^{d_s \times d}$ is a learnable matrix. Then the query (q), key (k), and value (v) can be computed as follows:

$$q_i = f_i^x W_q^s, k_j = f_j^x W_k^s, v_j = f_j^x W_v^s, \tag{7}$$

where W_q^s, W_k^s, and $W_v^s \in \mathbb{R}^{d \times d}$ are learnable projection matrices. Then the update of features f_i^x can be formulated as follows:

$$f_i^x \leftarrow f_i^x + (q_i \oplus \sum_j a_{ij} v_j) W_o^s, \tag{8}$$

where $W_o^s \in \mathbb{R}^{2d \times d}$ is a learnable projection matrix and $\oplus$ denotes the concatenation operator. The attention score $a_{ij} = \text{softmax}(q_i k_j^T / \sqrt{d})$.

GMSP Guided Cross-attention. In the cross-attention module, the query comes from one point cloud while the key and value come from the other. We design a GMSP-Guided Cross-attention module, aiming to strengthen the feature

interactions between semantically similar regions while suppressing the others. To achieve this, we formulate the feature interaction by exploiting the Gaussian Mixture Semantic Prior θ^x. Concretely, given a key f_i^x from point cloud $\dot{\mathbf{X}}$, θ_i^x acts as a semantic prior, where the weights indicate the semantic correlation between f_i^x and a certain region in point cloud $\dot{\mathbf{Y}}$. The attention score e_{ij} is defined as follows:

$$e_{ij} = \frac{f_i^x W_q^c (f_j^y W_k^c)^T \theta_{ij}^x}{\sqrt{d_t}}, \tag{9}$$

where W_q^c and $W_k^c \in \mathbb{R}^{d\times d}$ are learnable projection matrices. Then the cross-attention update of f_i^x can be formulated as follows:

$$f_i^x \leftarrow f_i^x + (f_i^x \oplus \sum_j a_{ij} f_j^y) W_o^c, \tag{10}$$

where $W_o^c \in \mathbb{R}^{2d\times d}$ is a learnable matrix. The attention score $a_{ij} = \text{softmax}(e_{ij})$.

4.4 Semantics Constrained Matching

Reliable correspondences exist between points or regions that share similar semantic information. To leverage this, we design a Semantic Constrained Matching module to perform feature matching only within the region sharing high semantic similarity. It enables us to eliminate the potential erroneous matching between semantically unrelated regions. To achieve this, we first compute the point-wise matching score:

$$s_{ij} = \frac{1}{\sqrt{d}} \langle f_i^x W_m, f_j^y W_m \rangle, \tag{11}$$

where $W_m \in \mathbb{R}^{d\times d}$ is a learnable projection matrix. Then, we determine whether a pair of points lies in the same semantic region by exploiting the mutual Gaussian Mixture Semantic Prior, i.e., θ_{ij}^x and θ_{ji}^y. Specifically, we apply the dual-softmax [29] on s_{ij}, then we set the matching score $C_{i,j}$ of a pair of points x_i and y_i to be zero if both θ_{ij}^x and θ_{ji}^y are under a predefined threshold τ_m:

$$C_{i,j} = \frac{s_{i,j}}{\sum_{k=1}^{|\dot{\mathbf{X}}|} s_{i,k}} \cdot \frac{s_{i,j}}{\sum_{k=1}^{|\dot{\mathbf{Y}}|} s_{k,j}} \cdot [\![\max(\theta_{ij}^x, \theta_{ji}^y) > \tau_m]\!], \tag{12}$$

where $[\![\cdot]\!]$ is the Iverson bracket.

4.5 Supervision Loss

Semantics Aware Focal Loss. We design a Semantics Aware Focal Loss by considering the following: 1) our matching process should exclusively operate on points situated within identical semantic regions; therefore, the supervision of point pairs outside of those regions is unnecessary. 2) We aim to enhance the feature interactions among similar semantic regions. To achieve this, we construct a set of salient semantic correspondences Ω^L similar to Eq. (4):

$$\Omega^L = \{(i,j) \mid \langle s_i^x, s_j^y \rangle > \eta_l, s_i^x \in S^x, s_j^y \in S^y\}, \tag{13}$$

where η_l is a similarity threshold. During the training phase, we align the two point clouds with ground truth transformation and obtain a set of ground truth matches κ_{gt} by selecting the point pairs lower than a predefined threshold. The rest of the point pairs form a set of negative matches Π_{gt}. Our loss is therefore defined as $L = L_P + L_N$, where L_P and L_N aim to maximize and minimize confidence scores of positive and negative pairs respectively:

$$L_P = -\frac{1}{|P|} \sum_{(i,j)\in P} \alpha(1 - \mathcal{C}(i,j))^\gamma \log \mathcal{C}(i,j), \tag{14}$$

$$L_N = -\frac{1}{|N|} \sum_{(i,j)\in N} \alpha \mathcal{C}(i,j)^\gamma \log(1 - \mathcal{C}(i,j)), \tag{15}$$

where $\mathcal{C}(i,j)$ is the confidence matrix from Eq. (12) and $P \in \kappa_{gt} \cap \Omega^L$ and $N \in \Pi_{gt} \cap \Omega^L$. α and γ are empirically set to 0.25 and 2, respectively.

5 Experimental Results

5.1 Datasets

3DMatch dataset. The 3DMatch dataset [50] is a widely used RGB-D indoor scene dataset designed for point cloud registration tasks. It consists of 46 scenes for training, 8 for validation, and 8 for testing. Following the approach of Predator [17], the test set is further subdivided into two evaluation subsets based on their overlapping ratio: 3DMatch and 3DLoMatch. Specifically, scene pairs with an extremely low overlapping ratio (i.e., less than 10%) are excluded. Those with an overlapping ratio below 30% are classified as 3DLoMatch, while the remaining pairs are included in the 3DMatch evaluation set.

KITTI dataset. The KITTI odometry dataset [13] comprises 11 LiDAR-scanned sequences of outdoor driving scenes with corresponding RGB frames. Following the protocols of [6, 17], we divide the sequences into three segments: 0-5 for training, 6-7 for validation, and 8-10 for testing. According to [3, 6], we only evaluate point cloud pairs within a distance of 10 meters from each other. The results are shown in Section 4.1 in the supplementary document.

5.2 Results

Evaluation metrics. Following [17,20], our method is evaluated on five metrics: 1) Inlier Ratio (IR), 2) Feature Matching Recall (FMR), 3) Registration Recall (RR), 4) Relative Rotation Error (RRE), and 5) Relative Translation Error (RTE). IR measures the fraction of correspondences with a distance less than $0.1m$ under the ground-truth transformation. FMR estimates the fraction of the point cloud with an inlier ratio exceeding 5%. RR calculates the fraction of point

Table 1: Registration evaluation results on 3DMatch and 3DLoMatch, with the top and second-ranking results highlighted in bold and underlined, respectively.

Method	Reference	3DMatch			3DLoMatch		
		FMR	IR	**RR**	FMR	IR	**RR**
PerfectMatch [15]	CVPR 2019	94.7	36.0	81.5	61.9	11.4	36.6
FCGF [15]	ICCV 2019	95.2	56.9	88.2	60.9	21.4	45.8
D3Feat [3]	CVPR 2020	95.8	39.0	85.8	69.3	13.2	40.2
SpinNet [1]	CVPR 2021	97.6	47.5	88.6	75.3	20.5	59.8
Predator [17]	CVPR 2021	96.7	58.0	91.8	78.6	26.7	62.4
CoFiNet [48]	CVPR 2021	98.1	49.8	89.3	83.1	24.4	67.5
YOHO [40]	ACMMM 2022	98.2	64.4	90.8	79.4	25.9	65.2
PCR-CG [52]	ECCV 2022	97.4	-	89.4	80.4	-	66.3
GeoTransformer [27]	CVPR 2022	97.9	71.9	92.0	88.3	43.5	<u>75.0</u>
RegTr [20]	CVPR 2022	-	-	92.0	-	-	64.8
Lepard [20]	CVPR 2022	<u>98.3</u>	55.5	93.5	84.5	26.0	69.0
RoReg [41]	TPAMI 2023	98.2	<u>81.6</u>	92.9	82.1	39.6	70.3
RoITr [49]	CVPR 2023	98.0	**82.6**	91.9	<u>89.6</u>	**54.3**	74.7
SIRA-PCR [5]	ICCV 2023	98.2	70.8	<u>93.6</u>	88.8	43.3	73.5
Point-TTA + DGR [16]	ICCV 2023	-	-	92.5	-	-	50.7
Point-TTA + PointDSC [16]	ICCV 2023	-	-	93.5	-	-	57.8
SemReg(ours)	-	**98.6**	56.2	**94.7**	**89.8**	<u>29.4</u>	**75.9**

cloud pairs with a transformation RMSE less than $0.2m$, indicating the quality of the final alignment. RRE measures the rotational error in degrees and RTE measures the translation error in centimeters. See supplementary for details.

Registration results. Registration recall (RR) is the primary indicator of registration performance as it directly reflects the quality of the final registration results. Following [17], we report the registration recall results. Concretely, we run 50K RANSAC iterations on the established correspondences to estimate the final transformation. Notably, similar to Lepard [20], our approach is a coarse registration approach, where our correspondences are established at the coarse points. Specifically, the number of our established correspondences is at the scale of $0.1K$-$1K$. In contrast, the number of the established dense correspondences by other methods, such as GeoTransfromer [27] and CoFiNet [48], is at the scale of $6K$-$7K$. As shown in Table 1, our method achieves an outstanding RR performance in both 3DMatch and 3DLoMatch evaluation sets, reaching 94.7% and 75.9%, respectively. Comparing to the recent state-of-the-art methods, our approach outperforms SIRA-PCR [5] by 1.1% and Lepard [20] by 1.2% on 3DMatch. We also achieve superior performance on 3DLoMatch, surpassing GeoTransformer [27] by 0.9% and RoReg [41] by 1.2%. Visualizations of registration samples are shown in Figure 5. We note that the coarse matching strategy results in more outliers under the same inlier threshold, as compared to a fine-grained approach such as GeoTransformer. However, borderline correspondences (outliers above but near the threshold) can also benefit successful registration [20]. As a result, we observe that a higher IR does not necessarily guarantee a higher RR. Please refer to the supplementary for detailed analysis.

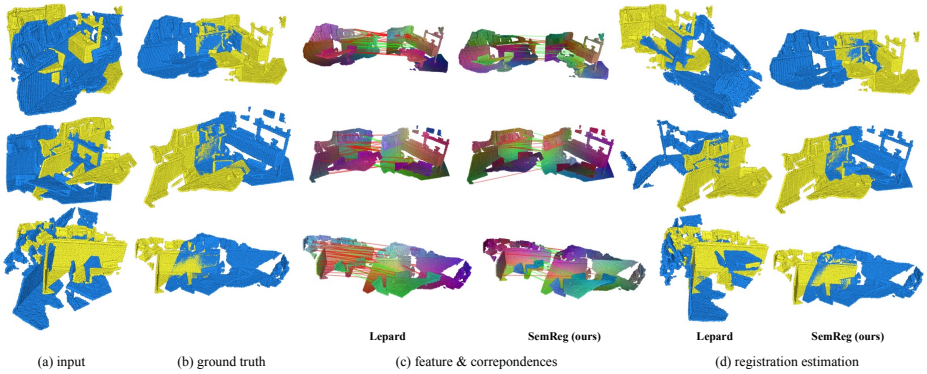

Fig. 5: Visualization comparisons between Lepard [20] and our proposed SemReg. Additional visualizations are shown in the supplementary.

Correspondence results. As shown in Table 1, despite the FMR saturating near 99%, our method still outperforms all state-of-the-art methods consistently, achieving a superior performance of 98.6% on the 3DMatch. This advantage is also evident on the 3DLoMatch, where our method achieves the leading result of 89.8%, surpassing SIRA-PCR [5] by 1.0% and Lepard [20] by 5.3%.

Table 2: Fine alignment evaluation results on 3DMatch and 3DLoMatch, with the top and second-ranking results highlighted in bold and underlined, respectively.

Method	3DMatch			3DLoMatch		
	RRE	RTE	RR	RRE	RTE	RR
RANSAC						
Predator [17]	2.72	7.8	91.8	4.44	11.6	62.4
GeoTransformer [27]	<u>2.31</u>	7.7	92.0	<u>3.96</u>	**10.6**	<u>75.0</u>
Lepard [20]	2.48	<u>7.2</u>	<u>93.5</u>	4.10	<u>10.8</u>	69.0
SemReg(ours)	**2.30**	**6.7**	**94.7**	**3.94**	11.2	**75.9**
RANSAC+ICP						
Predator [17]	2.06	6.2	92.3	3.46	9.8	65.2
GeoTransformer [27]	<u>1.91</u>	7.2	91.8	<u>2.82</u>	<u>9.5</u>	<u>74.5</u>
Lepard [20]	1.96	**6.0**	<u>93.9</u>	3.17	**8.9**	71.3
SemReg(ours)	**1.83**	<u>6.1</u>	**95.4**	**2.71**	<u>9.5</u>	**77.4**

Fine alignment results. Our proposed SemReg is characterized as a coarse registration approach, indicating a higher potential for fine alignment. To demonstrate this fine alignment capability, we follow Lepard [20] and conduct additional experiments. We use the final RANSAC estimated transformation as the initial pose and subsequently refine the alignment using the point-to-point ICP algorithm. We compare the RTE, RRE, and RR on both 3DMatch and 3DLoMatch evaluation sets. Please note that we obtain the results of GeoTransformer [27] by

Table 3: Ablation results showcasing the impact of Semantics Constrained Matching (SCM), Semantics Aware Focal Loss (SAFL), Semantics Embedded Self-attention (SESA), and GMSP Guided Cross-attention (GGCA) on the registration task. The "-" sign represents the corresponding module being removed.

SCM	SAFL	SESA	GGCA	3DMatch			3DLoMatch		
				FMR	IR	RR	FMR	IR	RR
-	-	-	-	97.6	45.3	92.1	82.3	22.6	66.5
✓	-	-	-	97.3	47.4	92.3	85.7	24.2	66.9
✓	✓	-	-	97.6	50.1	93.0	86.2	26.4	72.0
✓	✓	✓	-	98.5	53.4	93.7	86.6	27.3	73.7
✓	✓	-	✓	98.6	55.2	94.0	87.3	28.6	74.8
✓	-	✓	✓	98.0	54.6	94.3	87.7	28.8	75.1
-	✓	✓	✓	98.2	53.7	94.1	86.9	28.1	74.7
✓	✓	✓	✓	**98.6**	**56.2**	**94.7**	**88.4**	**29.4**	**75.9**

re-running the provided code along with the pre-trained weights, as they did not provide these metrics in their original paper. As shown in Table 2, our proposed SemReg achieves a remarkable RR improvement from 94.7% to 95.4% and from 75.9% to 77.4% for 3DMatch and 3DLoMatch, respectively. After the ICP refinement, our method outperforms the second-best methods Lepard [20] by 1.5% and GeoTransformer [27] by 2.9% for 3DMatch and 3DLoMatch, respectively. Our proposed SemReg also showcases competitive performance on the other two metrics. For RRE, our method achieves the top results consistently both before and after the ICP refinement. It also achieves the second-best results on the RTE, after the ICP refinement.

5.3 Ablation study

We conduct extensive ablation experiments to study the significance of each component in our proposed SemReg. We perform ablations on the four modules that we proposed: 1) Semantics Constrained Matching (SCM), 2) Semantics Aware Focal Loss (SAFL), 3) Semantics Embedded Self-attention (SESA), and 4) GMSP Guided Cross-attention (GGCA). When ablating the SCM and the SAFL modules, we substituted them with the matching module and the focal loss used in Lepard [20], respectively. Similarly, for the SESA and the GGCA modules, we substitute them with a vanilla attention module. The results are listed in Table 3. In general, when all our modules are removed from the pipeline (row 1), the RR drops dramatically from 94.7% to 92.1% and 75.9% to 66.5% for 3DMatch and 3DLoMatch, respectively. We summarize the ablation study by answering the following questions:

How effective are the SESA and GGCA modules? Registration performance saw a significant drop when the SESA and GGCA modules were both removed (row 3). Removing either the SESA or the GGCA module weakens the performance by over 0.7% (row 4 and row 5). Additionally, Figure 6 showcases the robustness of our GMSP under an extremely low overlap ratio, which enables GGCA to learn reliable feature representations.

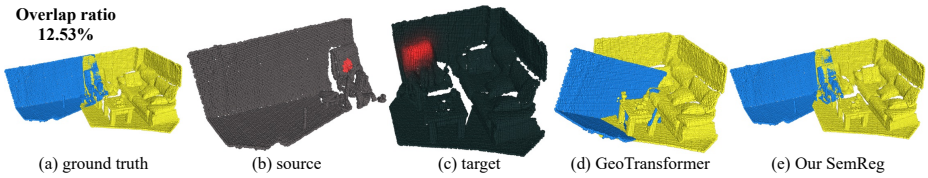

Fig. 6: An extremely low overlap case where only a chair is in the overlapping region. Given a query point from the source (red dot in (b)), our GMSP can still identify the semantically related region in the target (red area in (c)) and achieve successful registration (see (e)) while GeoTransformer yields a poor result (shown in (d)).

Is the SCM module significant? When only SCM is applied, the RR saw an improvement of 0.2% and 0.4% (row 2) for 3DMatch and 3DLoMatch, respectively. The removal of the SCM module exhibits a similar trend, resulting in a decrease of over 0.6% for RR (row 7).

Is the SAFL helpful? As shown in row 6, without the SAFL module, RR decreases by 0.4% for 3DMatch and 0.8% for 3DLoMatch.

These findings demonstrate the effectiveness and necessity of our proposed modules in boosting registration performance. Ablation experiments on the hyperparameters τ_m, η, and the number of SGFI blocks are in the supplementary.

6 Limitation

Our method is designed to perform registration on the point clouds derived from RGB-D images since it requires the aligned 2D image to extract semantic features. Consequently, our SemReg is cross-modal and does not directly operate on standalone point cloud data. However, in the absence of respective RGB frames, a promising solution is to first render 2D images from the point cloud data (see Section 4.2 in the supplementary document), which can be subsequently used to extract semantic features. We will further explore this in future work.

7 Conclusion

In this paper, we presented SemReg, a novel cross-modal point cloud registration framework that leverages semantic information inherent in 2D images to learn robust feature representations. Specifically, we design a Gaussian Mixture Semantic Prior that fuses 2D semantic features across RGB frames to reveal semantic correlations between regions across the point cloud pair. Subsequently, we propose the Semantics Guided Feature Interaction module that uses this prior to emphasize the feature interactions between the semantically similar regions during the cross-attention stage. In addition, we design a Semantics Aware Focal Loss that facilitates the learning of robust features, and a Semantics Constrained Matching module that performs matching only between regions sharing similar semantics. We conduct experiments on both indoor and outdoor datasets, and experimental results show the superiority of our method.

Acknowledgements

This work is supported in part by the project (No. 3.6267.01).

References

1. Ao, S., Hu, Q., Yang, B., Markham, A., Guo, Y.: Spinnet: Learning a general surface descriptor for 3d point cloud registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11753–11762 (2021)
2. Aoki, Y., Goforth, H., Srivatsan, R.A., Lucey, S.: Pointnetlk: Robust & efficient point cloud registration using pointnet. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7163–7172 (2019)
3. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6359–6367 (2020)
4. 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. Spie (1992)
5. Chen, S., Xu, H., Li, R., Liu, G., Fu, C.W., Liu, S.: Sira-pcr: Sim-to-real adaptation for 3d point cloud registration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14394–14405 (October 2023)
6. Choy, C., Park, J., Koltun, V.: Fully convolutional geometric features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8958–8966 (2019)
7. Deng, H., Birdal, T., Ilic, S.: Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In: Proceedings of the European conference on computer vision (ECCV). pp. 602–618 (2018)
8. Deng, H., Birdal, T., Ilic, S.: Ppfnet: Global context aware local features for robust 3d point matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 195–205 (2018)
9. Drost, B., Ulrich, M., Navab, N., Ilic, S.: Model globally, match locally: Efficient and robust 3d object recognition. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 998–1005. Ieee (2010)
10. El Banani, M., Johnson, J.: Bootstrap your own correspondences. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6433–6442 (2021)
11. Fung, S., Lu, X., Mykolaitis, M., Razzak, I., Kostkevičius, G., Ozerenskis, D.: Anatomical landmarks localization for 3d foot point clouds. In: International Conference on Neural Information Processing. pp. 627–638. Springer (2022)
12. Fung, S., Pan, W., Liu, X., Yearwood, J., Dazeley, R., Lu, X.: Topformer: Topology-aware transformer for point cloud registration. In: International Conference on Computational Visual Media. pp. 112–128. Springer (2024)
13. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
14. Georgakis, G., Karanam, S., Wu, Z., Ernst, J., Košecká, J.: End-to-end learning of keypoint detector and descriptor for pose invariant 3d matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1965–1973 (2018)

15. Gojcic, Z., Zhou, C., Wegner, J.D., Wieser, A.: The perfect match: 3d point cloud matching with smoothed densities. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5545–5554 (2019)
16. Hatem, A., Qian, Y., Wang, Y.: Point-tta: Test-time adaptation for point cloud registration using multitask meta-auxiliary learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16494–16504 (October 2023)
17. Huang, S., Gojcic, Z., Usvyatsov, M., Wieser, A., Schindler, K.: Predator: Registration of 3d point clouds with low overlap. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 4267–4276 (2021)
18. Huang, X., Mei, G., Zhang, J.: 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. pp. 11366–11374 (2020)
19. Lee, J., Kim, S., Cho, M., Park, J.: Deep hough voting for robust global registration. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15994–16003 (2021)
20. Li, Y., Harada, T.: Lepard: Learning partial point cloud matching in rigid and deformable scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5554–5564 (2022)
21. Lu, W., Wan, G., Zhou, Y., Fu, X., Yuan, P., Song, S.: Deepvcp: An end-to-end deep neural network for point cloud registration. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12–21 (2019)
22. Lu, X., Chen, H., Yeung, S.K., Deng, Z., Chen, W.: Unsupervised articulated skeleton extraction from point set sequences captured by a single depth camera. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
23. Lu, X., Deng, Z., Luo, J., Chen, W., Yeung, S.K., He, Y.: 3d articulated skeleton extraction using a single consumer-grade depth camera. Computer Vision and Image Understanding **188**, 102792 (2019)
24. Mei, G., Tang, H., Huang, X., Wang, W., Liu, J., Zhang, J., Van Gool, L., Wu, Q.: Unsupervised deep probabilistic approach for partial point cloud registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13611–13620 (2023)
25. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2906–2917 (2021)
26. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
27. Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Xu, K.: Geometric transformer for fast and robust point cloud registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11143–11152 (2022)
28. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
29. Rocco, I., Cimpoi, M., Arandjelović, R., Torii, A., Pajdla, T., Sivic, J.: Neighbourhood consensus networks. Advances in neural information processing systems **31** (2018)
30. Rusu, R.B., Blodow, N., Beetz, M.: Fast point feature histograms (fpfh) for 3d registration. In: 2009 IEEE international conference on robotics and automation. pp. 3212–3217. IEEE (2009)

31. Schnabel, R., Wahl, R., Klein, R.: Efficient ransac for point-cloud shape detection. In: Computer graphics forum. vol. 26, pp. 214–226. Wiley Online Library (2007)
32. Shao, D., Lu, X., Liu, X.: 3d intracranial aneurysm classification and segmentation via unsupervised dual-branch learning. IEEE Journal of Biomedical and Health Informatics **27**(4), 1770–1779 (2022)
33. Shao, D., Lu, X., Liu, X., Razzak, I.: Contrastive learning with self-reconstruction for 3d intracranial aneurysm detection. Available at SSRN 4405529 (2023)
34. Shen, Y., Hui, L., Jiang, H., Xie, J., Yang, J.: Reliable inlier evaluation for unsupervised point cloud registration. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2198–2206 (2022)
35. de Silva Edirimuni, D., Lu, X., Li, G., Wei, L., Robles-Kelly, A., Li, H.: Straightpcf: Straight point cloud filtering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20721–20730 (June 2024)
36. de Silva Edirimuni, D., Lu, X., Shao, Z., Li, G., Robles-Kelly, A., He, Y.: Iterativepfn: True iterative point cloud filtering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13530–13539 (June 2023)
37. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6411–6420 (2019)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
39. Wang, B., Chen, C., Cui, Z., Qin, J., Lu, C.X., Yu, Z., Zhao, P., Dong, Z., Zhu, F., Trigoni, N., et al.: P2-net: Joint description and detection of local features for pixel and point matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16004–16013 (2021)
40. Wang, H., Liu, Y., Dong, Z., Wang, W.: You only hypothesize once: Point cloud registration with rotation-equivariant descriptors. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 1630–1641 (2022)
41. Wang, H., Liu, Y., Hu, Q., Wang, B., Chen, J., Dong, Z., Guo, Y., Wang, W., Yang, B.: Roreg: Pairwise point cloud registration with oriented descriptors and local rotations. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
42. Wang, Y., Solomon, J.M.: Deep closest point: Learning representations for point cloud registration. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3523–3532 (2019)
43. Wang, Y., Solomon, J.M.: Prnet: Self-supervised learning for partial-to-partial registration. Advances in neural information processing systems **32** (2019)
44. Xu, H., Liu, S., Wang, G., Liu, G., Zeng, B.: Omnet: Learning overlapping mask for partial-to-partial point cloud registration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3132–3141 (2021)
45. Yew, Z.J., Lee, G.H.: 3dfeat-net: Weakly supervised local 3d features for point cloud registration. In: Proceedings of the European conference on computer vision (ECCV). pp. 607–623 (2018)
46. Yew, Z.J., Lee, G.H.: Rpm-net: Robust point matching using learned features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11824–11833 (2020)
47. Yew, Z.J., Lee, G.H.: Regtr: End-to-end point cloud correspondences with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6677–6686 (2022)

48. Yu, H., Li, F., Saleh, M., Busam, B., Ilic, S.: Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. Advances in Neural Information Processing Systems **34**, 23872–23884 (2021)
49. Yu, H., Qin, Z., Hou, J., Saleh, M., Li, D., Busam, B., Ilic, S.: Rotation-invariant transformer for point cloud matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5384–5393 (2023)
50. Zeng, A., Song, S., Nießner, M., Fisher, M., Xiao, J., Funkhouser, T.: 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1802–1811 (2017)
51. Zhang, D., Lu, X., Qin, H., He, Y.: Pointfilter: Point cloud filtering via encoder-decoder modeling. IEEE Transactions on Visualization and Computer Graphics **27**, 2015–2027 (2021)
52. Zhang, Y., Yu, J., Huang, X., Zhou, W., Hou, J.: Pcr-cg: Point cloud registration via deep explicit color and geometry. In: European Conference on Computer Vision. pp. 443–459. Springer (2022)
53. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE multimedia **19**(2), 4–10 (2012)
54. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16259–16268 (2021)