Appendix

A. Definitions of Equivariance and Invariance

For a specified function $f : X \to Y$ as well as a specified group action G, f is said to be equivariant with respect to transformation action $g \in G$ if,

$$f(g \circ x) = g \circ f(x), \quad x \in X \tag{6}$$

Analogously, f is said to be invariant to transformations $g \in G$ when the following equation is satisfied:

$$f(g \circ x) = f(x), \quad x \in X \tag{7}$$

B. Theoretical Proof of Equivariance

Lemma 1. *Given a discrete 2D rotation group*[†] $\mathcal{R} \subset \mathbf{SO}(2)$*, where* $\mathcal{R} = \{r_i \in \mathbb{R}^{3\times3}, i = 1, 2, ..., L\}$*, then the proposed spatial point transformer is an equivariant map for the 2D rotation group* $\mathcal{R}$*.*

Proof: For a local patch $\mathbf{P}^{\mathrm{s}}$, the spatial point transformer in our framework can be regarded as a mapping $\mathcal{M}_v$ from $\mathbf{P}^{\mathrm{s}}$ to cylindrical volume $\mathbf{C} \in \mathbb{R}^{J\times K\times L\times k_v\times 3} : \mathbb{R}^{3\times|\mathbf{P}^{\mathrm{s}}|} \to \mathbb{R}^{J\times K\times L\times k_v\times 3}$. For a group action r_i in $\mathcal{R}$, suppose $\widetilde{\mathbf{P}}^{\mathrm{s}} = r_i \circ \mathbf{P}^{\mathrm{s}} = r_i\mathbf{P}^{\mathrm{s}}$, and the rotated local neighbouring set $\widetilde{\mathbf{P}}_{jk(l+i)} = r_i\mathbf{P}_{jkl}$. On the other hand, for the rotation matrix defined in Eq. 3, we have $\mathbf{R}_{jkl} = \mathbf{R}_{jk(l+i)}112r_i$. Then, the (j^{th}, k^{th}, l^{th}) element $\mathbf{c}^p_{jkl}$ of cylindrical volume $\mathbf{C}$ satisfies:

$$\begin{aligned}\mathbf{c}^p_{jkl} &= \mathbf{R}_{jkl}\mathbf{P}_{jkl} = \mathbf{R}_{jk(l+i)}r_i\mathbf{P}_{jkl} \\ &= \mathbf{R}_{jk(l+i)}\widetilde{\mathbf{P}}_{jk(l+i)} = \mathbf{c}^{\widetilde{p}}_{jk(l+i)},\end{aligned} \tag{8}$$

where $\mathbf{c}^{\widetilde{p}}_{jk(l+i)} \in \widetilde{\mathbf{C}}$, which is the cylindrical volume corresponding to the $\widetilde{\mathbf{P}}^{\mathrm{s}}$. Based on Eq. 8, we can infer that $\mathbf{c}^p_{jk(l-i)} = \mathbf{c}^{\widetilde{p}}_{jkl}$, hence the transformed cylindrical volume $\widetilde{\mathbf{C}}$ can be formulated as:

$$\begin{aligned}\widetilde{\mathbf{C}} =& \mathcal{M}_v(r_i \circ \mathbf{P}^{\mathrm{s}}) = \mathcal{M}_v(\widetilde{\mathbf{P}}^{\mathrm{s}}) \\ &= [\mathbf{c}^{\widetilde{p}}_{111}, ..., \mathbf{c}^{\widetilde{p}}_{jkl}, ..., \mathbf{c}^{\widetilde{p}}_{JKL}] \\ &= [\mathbf{c}^p_{11(1-i)}, ..., \mathbf{c}^p_{jk(l-i)}, ..., \mathbf{c}^p_{JK(L-i)}],\end{aligned} \tag{9}$$

where $\mathbf{c}^p_{jkl} = \mathbf{c}^p_{jk(l+L)}$ if $l < 1$, due to the periodic property of the cylindrical volume in the XY plane. On the other hand, $r_i \circ \mathcal{M}_v$ means rotating the cylindrical volume $\mathbf{C}$ around the Z-axis, that is:

$$\begin{aligned}&r_i \circ \mathcal{M}_v(\mathbf{P}^{\mathrm{s}}) = r_i \circ [\mathbf{c}^p_{111}, ..., \mathbf{c}^p_{jkl}, ..., \mathbf{c}^p_{JKL}] \\ &= [\mathbf{c}^p_{11(1-i)}, ..., \mathbf{c}^p_{jk(l-i)}, ..., \mathbf{c}^p_{JK(L-i)}] \\ &= \mathcal{M}_v(r_i \circ \mathbf{P}^{\mathrm{s}}),\end{aligned} \tag{10}$$

which completes our proof that the spatial point transformer $\mathcal{M}_v$ is an equivariant map for the rotation group $\mathcal{R}$.

Lemma 2. *Given a discrete 2D rotation group* $\mathcal{R} \subset \mathbf{SO}(2)$*, where* $\mathcal{R} = \{r_i \in \mathbb{R}^{3\times3}, i = 1, 2, ..., L\}$*, then 3DCCN is an equivariant map for the 2D rotation group* $\mathcal{R}$*.*

Proof: The proposed 3D cylindrical convolution can be formulated as a set of convolution filter ψ^i on the cylindrical feature maps f:

$$\begin{aligned}&(f * \psi^i)(\rho, z, \theta) = \\ &\sum_d \sum_j \sum_k \sum_l f_d(j, k, l)\psi^i_d(j - \rho, k - z, l - \theta),\end{aligned} \tag{11}$$

where ρ, θ and z denote radial distance, azimuth angle and height, respectively. d is the number of channels in feature map.

Suppose a group action r_i in $\mathcal{R}$ operating on cylindrical feature maps f, we have $(r_i \circ f)(\rho, z, \theta) = f(\rho, z, \theta - i)$. To clarify, the r_i-transformed feature maps $r_i \circ f$ at the coordinate (ρ, z, θ) is equivalent to find the value in the original feature map f at the coordinate $(\rho, z, \theta - i)$. Leaving out the summation over feature maps for clarity, we have:

$$\begin{aligned}&((r_i \circ f) * \psi^i)(\rho, z, \theta) = \\ &\sum_j \sum_k \sum_l f(j, k, l - i)\psi^i(j - \rho, k - z, l - \theta).\end{aligned} \tag{12}$$

Using the substitution $l \to l + i$, then Eq. 12 can be transformed into:

$$\begin{aligned}&((r_i \circ f) * \psi^i)(\rho, z, \theta) \\ &= \sum_j \sum_k \sum_l f(j, k, l)\psi^i(j - \rho, k - z, l - (\theta - i)) \\ &= (f * \psi^i)(\rho, z, \theta - i) \\ &= (r_i \circ (f * \psi^i))(\rho, z, \theta),\end{aligned} \tag{13}$$

which completes our proof that 3DCCN is an equivariant map for the 2D rotation group $\mathcal{R}$.

C. Detailed Network Architecture

Using 3D Cylindrical Convolution (3D-CCN) as a basic operator, we build a hierarchical learning architecture as depicted in Figure 6. To ensure the reproducibility of our framework, we also provide detailed information on the kernel size, stride, and the number of filters in this figure. A number of cylindrical convolution layers are stacked together to progressively learn descriptive, yet compact local

[†]The minimum rotation unit depends on the way partion along the azimuth axis. *i.e.,* $\frac{2\pi}{L}$.

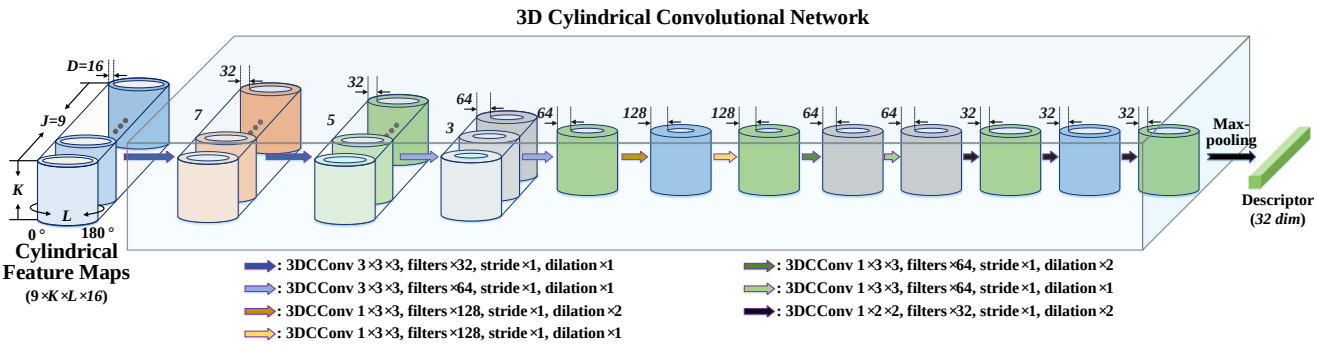

Figure 6: Detailed architecture of our proposed 3D cylindrical convolution networks.

feature representations. In particular, the maximum number of channels used in our cylindrical feature map is 128, which is much smaller than 1024 used in D3Feat [2]. This further makes our network very lightweight and less prone to overfitting.

D. Discussions of Equivariance and Invariance

In this paper, the invariant features, acquired by max-pooling the equivariant cylindrical features, are used for correspondence matching. Actually, combining invariant and equivariant features is an interesting idea to explore. Similar to [15], our SpinNet can be extended to direct pairwise registration with minor modifications. The point correspondences can be firstly estimated through invariant feature matching, and then the relative transformation can be calculated based on the discrepancy between equivariant features of each pair of point correspondence. On the other hand, we can also calculate a canonical orientation for each patch based on its equivariant feature similar to [51], thereby further estimating the relative transformation between two paired patches. In this case, each pair of correspondence correlates with a transformation hypothesis, hence the final transformation to align two point clouds can be readily obtained using clustering. Overall, compared with existing transformation calculation methods such as RANSAC, combining invariant and equivariant features can eliminate the combinatorial explosion of feature correspondences, but also improve the reliability of the estimated transformation [26].

E. Relating to Prior Works

Despite the resemblance of vocabulary, our SpinNet differs from SpinImages (SIs) [30] and 3D Shape Contexts (3DSC) [19] in several aspects: (1) **Rotation invariance**. Both SIs and 3DSC rely on the handcrafted point density, while our SpinNet explicitly transforms the point clouds into cylindrical volumes based on the spatial point transformer, enabling rotation invariance with end-to-end optimization. (2) **Descriptiveness**. Both SIs and 3DSC encode the local surface by simply counting the number of points that fall into each bin, while our SpinNet leverages the powerful neural feature extractor to learn local geometrical patterns from each spherical voxel and its local context. Hence, the learned local feature is descriptive and robust. (3) **Compactness**. Our descriptor is more compact (32 channels), compared with SIs (225) and 3DSC (1980) descriptors.

F. Detailed Evaluation Metrics

We further provide the detailed evaluation metrics used in our experiments (Sec. 4).

Evaluation Metrics on 3DMatch and ETH. We adopt Feature Matching Recall (FMR) as the main evaluation metric to evaluate the performance of the learned descriptors. Similar to [5, 14, 13, 8], we also provide a formal definition for each metric as follows.

First, suppose there are a total of H pairs of fragments in the 3DMatch dataset, where the overlap is greater than 30%. Each pair of fragments $\mathcal{P}_h$ and $\mathcal{Q}_h$ can be aligned by the ground-truth rigid transformation $\mathbf{T}_h = \{\mathbf{R}_h, \mathbf{t}_h\}$. Then, we randomly select n points from the two point clouds to obtain $\mathcal{P}_h^n = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n\}$ and $\mathcal{Q}_h^n = \{\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_n\}$. In particular, a set of point correspondences $\mathbf{\Omega}_h$ between $\mathcal{P}_h^n$ and $\mathcal{Q}_h^n$ is also generated by applying nearest neighbor search NN in the feature space $\mathcal{M}$:

Then the average feature matching recall on the 3DMatch dataset is defined as:

$$\text{FMR} = \frac{1}{H}\sum_{h=1}^{H}\mathbb{1}\left(\left[\frac{1}{|\mathbf{\Omega}_h|}\sum_{(\mathbf{p}_i,\mathbf{q}_j)\in\mathbf{\Omega}_h}\mathbb{1}(\|\mathbf{p}_i' - \mathbf{q}_j\| < \tau_1)\right] > \tau_2\right), \tag{14}$$

where $\mathbf{p}_i' = \mathbf{R}_h\mathbf{p}_i + \mathbf{t}_h$, $||\cdot||$ denotes the Euclidean distance, τ_1 and τ_2 is the inlier distance threshold and inliner ratio threshold, respectively. $\mathbb{1}$ is the indicator function. $\mathbf{\Omega}_h$

denotes a set of point correspondences between $\mathcal{P}_h^n$ and $\mathcal{Q}_h^n$. In particular, it is generated by applying nearest neighbor search NN in the feature space $\mathcal{M}$:

$$\begin{aligned}\mathbf{\Omega}_h = \{\{\mathbf{p}_i, \mathbf{q}_j\} | \mathcal{M}(\mathbf{p}_i) = \mathrm{NN}(\mathcal{M}(\mathbf{q}_j), \mathcal{M}(\mathcal{P}_h^n)), \\ \mathcal{M}(\mathbf{q}_j) = \mathrm{NN}(\mathcal{M}(\mathbf{p}_i), \mathcal{M}(\mathcal{Q}_h^n))\}.\end{aligned} \tag{15}$$

Evaluation Metrics on KITTI. Different from the indoor 3DMatch dataset, the evaluation metrics on the KITTI dataset are Relative Translational Error (RTE), Relative Rotation Error (RRE), and Success Rate (SR), respectively. According to the definitions in [39, 60, 8], for a pair of fragments $\mathcal{P}_h$ and $\mathcal{Q}_h$, the relative rotation error RRE is calculated as:

$$\mathrm{RRE} = \arccos\left(\frac{\mathrm{trace}(\hat{\mathbf{R}}_h^T \mathbf{R}_h) - 1}{2}\right)\frac{180}{\pi}, \tag{16}$$

where $\mathbf{R}_h$ and $\hat{\mathbf{R}}_h$ denote the ground-truth and the estimated rotation matrix, respectively. Analogously, the relative translation error RTE can be calculated by:

$$\mathrm{RTE} = \left\|\hat{\mathbf{t}}_h - \mathbf{t}_h\right\|, \tag{17}$$

where $\mathbf{t}_h$ and $\hat{\mathbf{t}}_h$ denote the ground-truth and the estimated translation matrix, respectively. Finally, success rate SR is defined as:

$$\mathrm{SR} = \frac{1}{H}\sum_{h=1}^{H} \mathbb{1}\Big(\mathrm{RRE}_h < 2\mathrm{m} \;\&\&\; \mathrm{RTE}_h < 5^\circ\Big). \tag{18}$$

G. Implementation Details

Here we provide extra implementation details in this section. The detailed hyperparameter settings of our SpinNet on different datasets are listed in Table 8. In particular, we keep the same parameter settings as the training dataset when generalized to unseen datasets, except for the support radius R and query radius R_v, due to the varying point densities in different datasets. Specifically, we follow the scheme in D3Feat [2] to adaptively adjust the radius according to the ratio.

Dataset	J	K	L	R	R_v	k_v
3DMatch [65]	9	40	80	0.3m	0.04m	30
KITTI [20]	9	30	60	2.0m	0.30m	30
ETH [46]	9	40	80	0.8m	0.10m	30

Table 8: The hyperparameters set by our method in different datasets.

H. Varying Hyperparameters

We further evaluate the performance of our SpinNet under varying parameters. For clarity, we have conducted preliminary experiments by arbitrarily combining the varying J, K, L ($J = \{7, 9, 11\}$, $K = \{40, 60, 80\}$, $L = \{20, 30, 40\}$) and report the quantitative results on the 3DMatch dataset in Table 9. It can be seen that the performance difference in FMR is less than 0.5%. This means that SpinNet is robust and not sensitive to varying hyperparameters.

J	K	L	R	R_v	k_v	FMR (%)	STD
7	20	40	0.3m	0.04m	30	97.8	2.0
7	30	60	0.3m	0.04m	30	97.5	1.8
7	40	80	0.3m	0.04m	30	97.4	1.8
9	20	40	0.3m	0.04m	30	97.6	1.9
9	30	60	0.3m	0.04m	30	97.7	1.9
9	40	80	0.3m	0.04m	30	97.6	1.9
11	20	40	0.3m	0.04m	30	97.5	1.8
11	30	60	0.3m	0.04m	30	97.5	1.8
11	40	80	0.3m	0.04m	30	97.6	1.9

Table 9: Quantitative results of our method under varying hyperparameters on the 3DMatch dataset.

I. Rotation Augmentation and Distribution

According to the results in Table 10 and Table 11, it can be seen that FCGF [8] and D3Feat [2] are robust again rotation. Fundamentally, the good performance of D3Feat and FCGF on the rotated 3DMatch dataset relies on extensive rotation-based data augmentation. However, the D3Feat and FCGF models trained on the rotated 3DMatch cannot generalize to ETH and KITTI dataset, because these three datasets (3DMatch, ETH, and KITTI) have significantly different rotation distributions. For illustration, we show the differences of rotation distributions in Fig. 7. In particular, the KITTI and ETH datasets only have SO(2) rotations (Fig. 7b), while the original and rotated 3DMatch datasets have different SO(3) rotations. However, for these three datasets, our SpinNet can well generalize across them all without relying on any data augmentation, demonstrating that our framework is indeed effective to overcome the rotation variance.

J. Additional Results on 3DMatch

For comparison, we also report the detailed quantitative results of our SpinNet on the 3DMatch dataset in Table 10 and the rotated 3DMatch dataset in Table 11.

	FPFH [48]	SHOT [55]	3DMatch [65]	CGF† [32]	PPFNet [14]	PPF-FoldNet [13]	PerfectMatch [22]	FCGF [8]	D3Feat [2]	LMVD [37]	Ours
Kitchen	30.6	17.8	57.5	46.1	89.7	78.7	97.0	-	-	**99.4**	99.2
Home 1	58.3	37.2	73.7	61.5	55.8	76.3	95.5	-	-	**98.7**	98.1
Home 2	46.6	33.7	70.7	56.3	59.1	61.5	89.4	-	-	94.7	**96.2**
Hotel 1	26.1	20.8	57.1	44.7	58.0	68.1	96.5	-	-	**99.6**	**99.6**
Hotel 2	32.7	22.1	44.2	38.5	57.7	71.2	93.3	-	-	**100.0**	97.1
Hotel 3	50.0	38.9	63.0	59.3	61.1	94.4	98.2	-	-	**100.0**	**100.0**
Study	15.4	7.2	56.2	40.8	53.4	62.0	94.5	-	-	95.5	**95.6**
MIT Lab	27.3	13.0	54.6	35.1	63.6	62.3	93.5	-	-	92.2	**94.8**
Average	35.9	23.8	59.6	47.8	62.3	71.8	94.7	95.2	95.8	97.5	**97.6**
STD	13.4	10.9	8.8	9.4	10.8	10.5	2.7	2.9	2.9	2.8	**1.9**

Table 10: Average recall (%) of different methods on the 3DMatch benchmark with $\tau_1 = 10$cm and $\tau_2 = 0.05$. The symbol '-' means the results are unavailable and † means the results are reported from [13], which is different from Table 1.

	FPFH [48]	SHOT [55]	3DMatch [65]	CGF† [32]	PPFNet [14]	PPF-FoldNet [13]	PerfectMatch [22]	FCGF [8]	D3Feat [2]	LMVD [37]	Ours
Kitchen	29.1	17.8	0.4	44.7	0.2	78.9	97.2	-	-	-	**99.0**
Home 1	59.0	35.6	1.3	66.7	0.0	78.2	96.2	-	-	-	**98.7**
Home 2	47.1	33.7	3.4	52.9	1.4	64.4	90.9	-	-	-	**96.2**
Hotel 1	30.1	21.7	0.4	44.3	0.4	67.7	96.5	-	-	-	**99.6**
Hotel 2	30.0	24.0	0.0	44.2	0.0	62.9	92.3	-	-	-	**97.1**
Hotel 3	51.9	33.3	1.0	63.0	0.0	96.3	98.2	-	-	-	**100.0**
Study	15.8	8.2	0.0	41.8	0.0	62.7	94.5	-	-	-	**94.9**
MIT Lab	41.6	62.3	3.9	45.5	0.0	67.5	93.5	-	-	-	**94.8**
Average	36.4	23.4	1.1	49.9	0.3	73.1	94.9	95.3	95.5	96.9	**97.5**
STD	13.6	9.5	1.2	8.9	0.5	10.4	2.5	3.3	3.5	-	1.9

Table 11: Average recall (%) of different methods on the rotated 3DMatch benchmark with $\tau_1 = 10$cm and $\tau_2 = 0.05$. The symbol '-' means the results are unavailable and † means the results are reported from [13], which is different from Table 1.

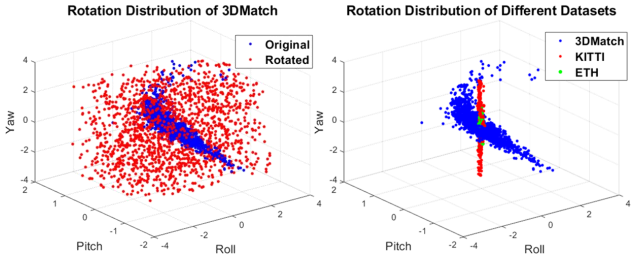

Figure 7: Rotation distribution of different datasets by plotting three Euler angles in each paired fragments.

K. Additional Qualitative Results.

As illustrated in Sec. 4.3, our SpinNet has demonstrated superior quantitative generalization performance across different datasets with different sensor modalities. Here, we further show additional qualitative results in this section.

Additional qualitative results on the 3DMatch dataset. We first show the additional qualitative results achieved by FCGF [8], D3Feat [2], and our SpinNet on the 3DMatch dataset in Fig. 8. It can be seen that the FCGF and D3Feat are prone to mismatching the fragments when the two input partial scans have relatively significant differences. However, our simple SpinNet can always achieve consistent registration performance on this dataset, despite only being trained on the outdoor KITTI dataset with sparse LiDAR point clouds.

Additional qualitative results on the KITTI dataset. Then, we show the extra qualitative results achieved by FCGF [8], D3Feat [2], and our SpinNet on the KITTI dataset in Fig. 9. We can clearly see that the point cloud in the KITTI dataset is significantly different from the point cloud in 3DMatch, since the KITTI dataset is mainly composed of *large-scale*, *sparse*, and *partial* LiDAR scans. As shown in Figure, FCGF and D3Feat tend to misalign the input fragments when the scene contains lots of geometrically-similar objects (*e.g.,* cars). However, our method can still achieve satisfactory registration results when only trained on the indoor 3DMatch dataset. This further validates the superior generalization ability of our method.

Additional qualitative results on the ETH dataset. We finally show the extra qualitative results achieved by FCGF [8], D3Feat [2], and our SpinNet on the ETH dataset in Fig. 10. Compared with the 3DMatch and KITTI data sets, the ETH dataset is collected by static terrestrial lasers in outdoor scenes, and is mainly composed of bushes and vegetation. As shown in Figure, it is highly challenging for FCGF and D3Feat to successfully align the input scans together, since this dataset suffers from issues such as noise, clutter, and occlusions. Nevertheless, the proposed SpinNet can still achieve excellent performance on this dataset.

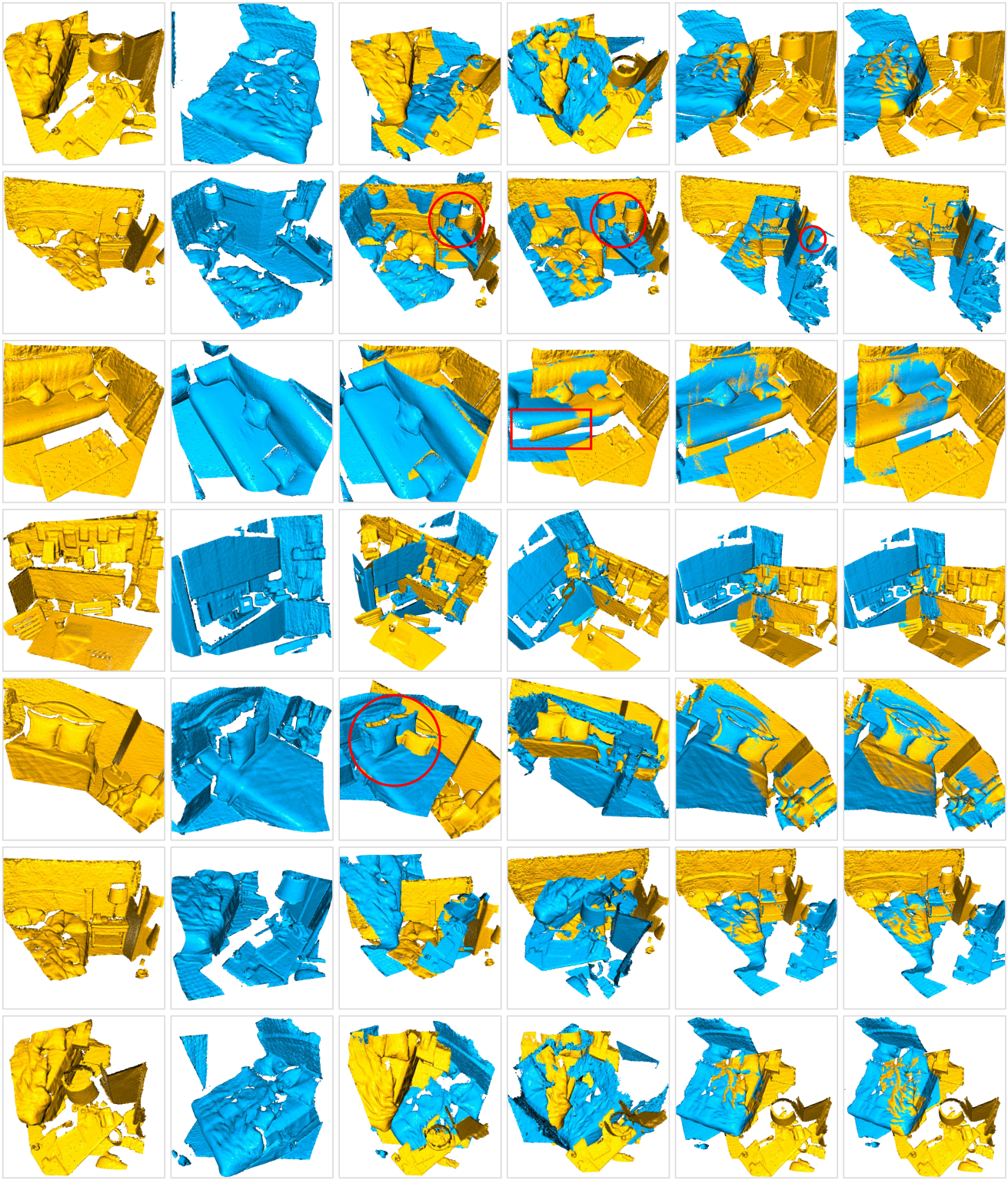

Figure 8: Additional qualitative results achieved by FCGF [8], D3Feat [2], and our **SpinNet** on the 3DMatch dataset. Note that, all methods are only trained on the outdoor KITTI [20] dataset. Red boxes/circles show the failure cases.

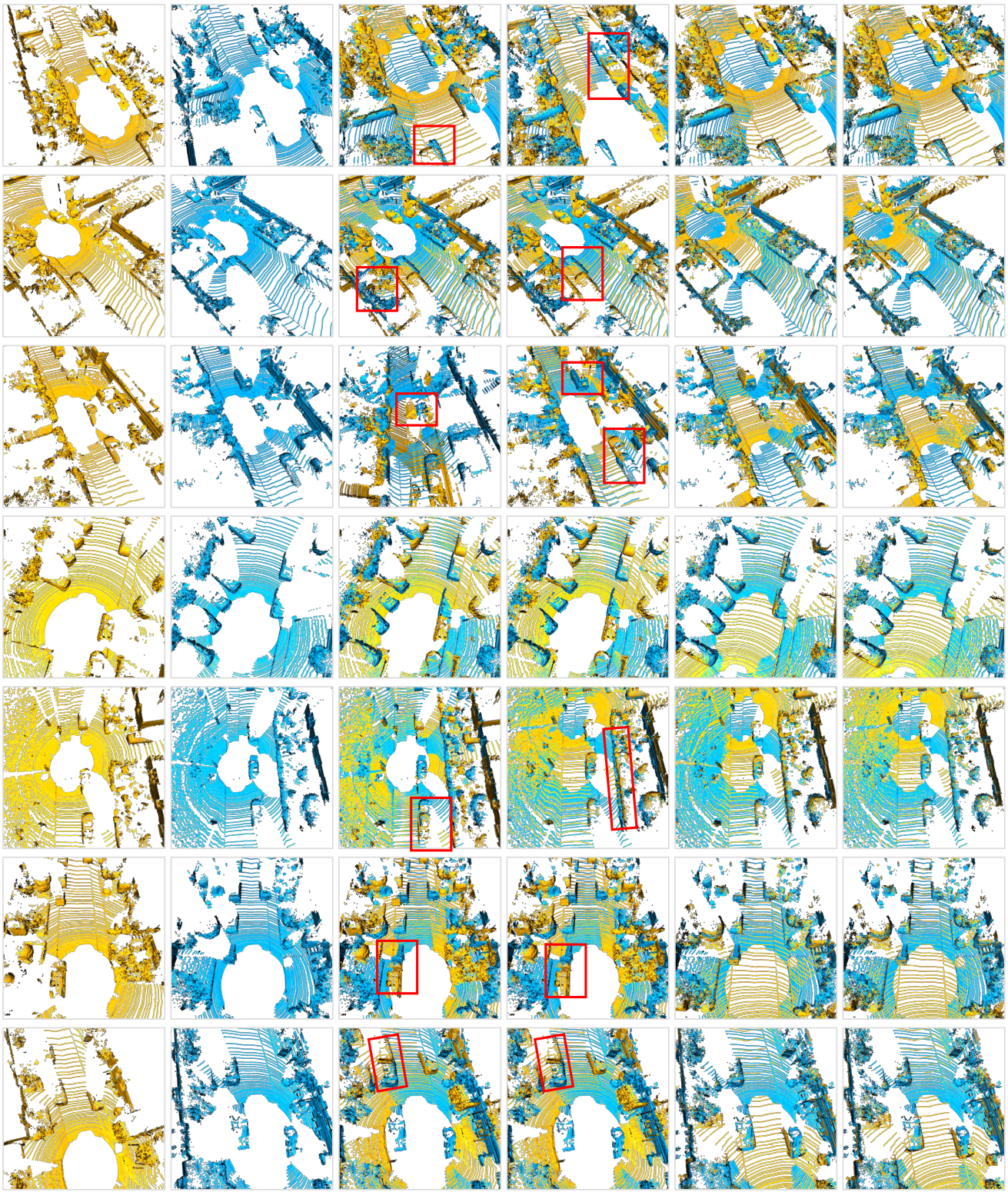

Figure 9: Additional qualitative results achieved by FCGF [8], D3Feat [2], and our **SpinNet** on the KITTI dataset. Note that, all methods are only trained on the indoor 3DMatch [65] dataset. Red boxes show the failure cases.

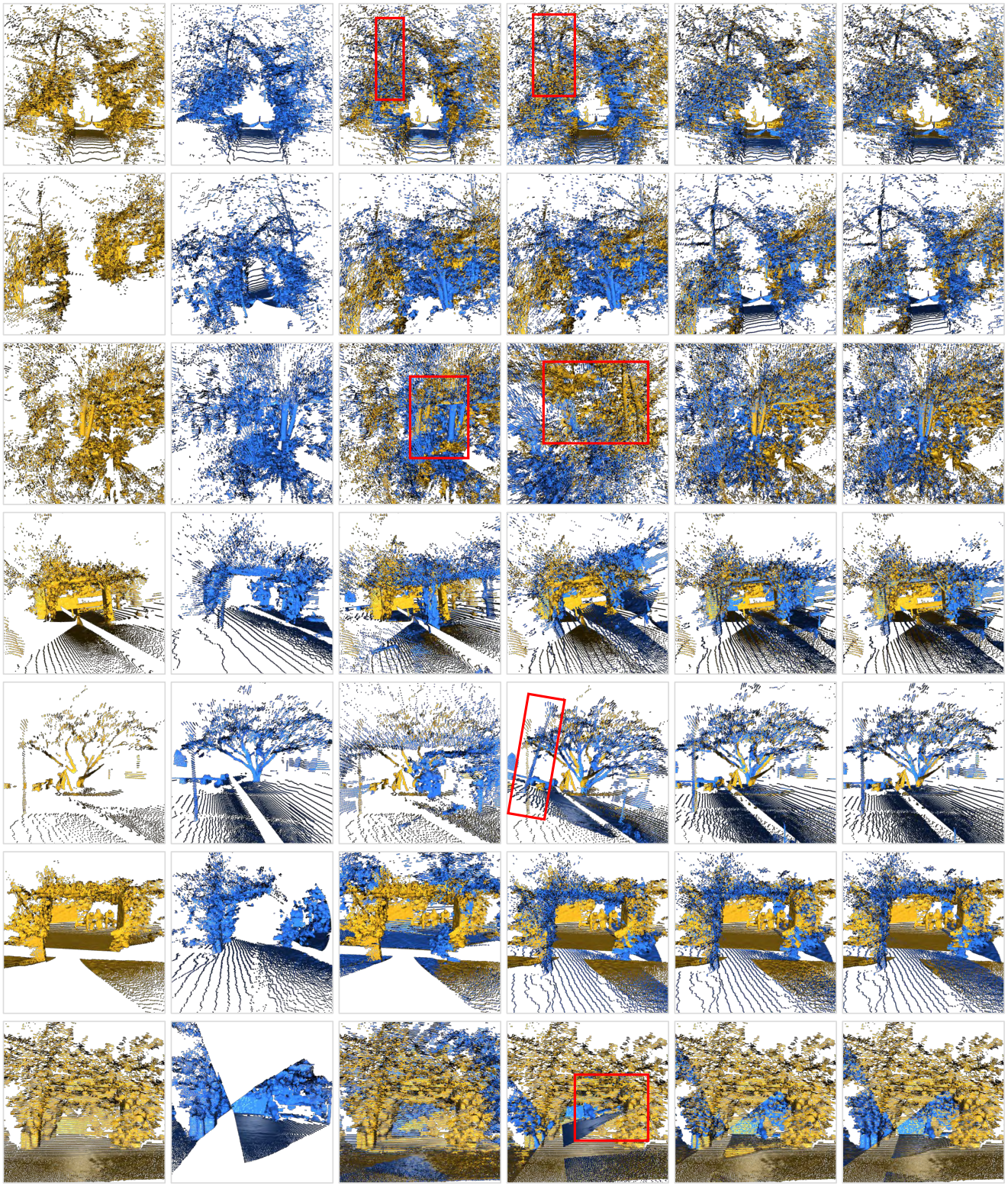

Figure 10: Additional qualitative results achieved by FCGF [8], D3Feat [2], and our **SpinNet** on the ETH dataset. Note that, all methods are only trained on the indoor 3DMatch [65] dataset. Red boxes/circles show the failure cases.