SG-PGM: Partial Graph Matching Network with Semantic Geometric Fusion for 3D Scene Graph Alignment and Its Downstream Tasks

Supplementary Material

Abstract

In the supplemental material, we provide additional details about the following:

- *Details on implementation. (Section A),*
- *Evaluation metrics of 3D scene graph alignment and downstream tasks (Secion B),*
- *Evaluation on scene graph alignment with controlled semantic noise and with predicted 3D scene graph (Section C),*
- *Additional ablation study on registration strategy and network variants (Section D),*
- *Visualisation on point cloud registration and point cloud mosaicking (Section E).*

A. Implementation Details

Data Generation for Alignment in Dynamics: To evaluate scene graph alignment in the changing environment Section 4.3, we generate the samples using the sub-scenes in the validation split and the original 3D scene maps from [41, 46]. The dynamics between scan and rescan of the same indoor scene consist of three types: ”non-rigid”, ”removed” and ”rigid”. We ignore small rigid object changes, whose Euler angles $\alpha + \beta + \gamma < 3°$, and mark them as aligned node ground truth. Thus, the sample numbers of scenarios (i), (ii) and (iii) are 819, 354, and 1,635.

Network and Training: We take the fine-level geometric feature of the KPConv-FPN as the input of our P2SG Fusion module. Same as suggested in [44], the input node embeddings of the AFA-U module are set to **zero vectors** for one graph and **one-hot vectors** for the other graph. Unlike in [44], we train the AFA-U module together with the other parts of the network in one stage. We employ the matching rescoring on the super-point matching stage of [31] because the fine-level points within a super-point are considered most likely to belong to the same object. The training procedure takes 10 epochs with the ADAM optimizer and an initial learning rate of $1e^{-4}$, which decreases by 0.1 every 4 epochs. If not specified, we mask out the unmatched objects of the scene fragments and conduct registration on the overlap region as a whole instead of registration traverse through all matched pairs.

B. Evaluation Metrics

We give the definition of evaluation metrics used in the main paper here. For the same evaluation metric used in multiple tasks, its definition will be adjusted based on input.

B.1. Scene Graph Alignment

Hits@K describes the fraction of true entities that appear in the first k entities of the sorted rank list R of the alignment prediction $\tilde{\mathbf{S}}$. Denoting the set of individual ranks as r_i, it is given as:

$$H_k(r_1, ..., r_n) = \frac{1}{n}\sum_{i}^{n}[r_i < k] \quad \in [0, 1] \tag{1}$$

where $[\cdot]$ is the Iversion bracket.

Mean Reciprocal Rank (MRR) is the arithmetic mean over the reciprocals of ranks of true triples:

$$MRR(r_1, ..., r_n) = \frac{1}{n}\sum_{i}^{n}\frac{1}{r_i} \quad \in (0, 1] \tag{2}$$

F1-score is the harmonic mean of the precision and recall. More specifically, the F1 score for graph matching is defined as:

$$\begin{aligned} tp,\ fp,\ fn &= \tilde{\mathbf{S}}\mathbf{S},\ \tilde{\mathbf{S}}(1-\mathbf{S}),\ (1-\tilde{\mathbf{S}})\mathbf{S} \\ F1 &= \frac{2tp}{2tp + fp + fn} \quad \in [0, 1]\,. \end{aligned} \tag{3}$$

B.2. Overlap Checking

Overlap checking of two 3D scenes is a binary classification problem that checks whether two 3D scenes overlap or not. Metrics (Precision, Recall, and F1-score) are given as:

$$\begin{aligned} Prec. &= \frac{TP}{TP + FP} \quad \in [0, 1]\,, \\ Recall &= \frac{TP}{TP + FN} \quad \in [0, 1]\,, \\ F1 &= 2\frac{Prec. \times Recall}{Prec. + Recall} \quad \in [0, 1]\,, \end{aligned} \tag{4}$$

in which TP is true positive, FP is false positive and FN as false negative.

B.3. Point Cloud Registration

Registration Recall (RR) is the fraction of successfully registered point cloud pairs. A point cloud pair is successfully registered when its transformation error is lower than threshold $\tau_1 = 0.2m$. In addition, the transformation error is the root mean square error of the ground truth correspondence C, to which the estimated transformation $\tilde{\mathbf{T}}$ has applied:

$$\begin{aligned} RMSE &= \sqrt{\frac{1}{|C|}\sum_{(p_x, q_y)\in C}\left\|\tilde{\mathbf{T}}(\mathbf{p}_x) - \mathbf{q}_y\right\|_2^2}, \\ RR &= \frac{1}{M}\sum_{i=1}^{M}[RMSE < \tau_1] \quad \in [0, 1]\,, \end{aligned} \tag{5}$$

where p_x and q_y denote the x-th point in source P and y-th point in reference Q, respectively; $[\cdot]$ is the inerson bracket; and M is the number of all point cloud pairs.

Feature Matching Recall (FMR) is the fraction of point cloud pairs whose Inlier Ration (IR) is above $\tau_3 = 0.05$. FMR measures the potential success during the registration, while Inlier Ratio is the fraction of inlier correspondences among all hypothesized correspondences $\tilde{C}$:

$$
\begin{aligned}
IR &= \frac{1}{\left|\tilde{C}\right|} \sum_{(p_x, q_y)\in\tilde{C}} \left[\left\|\mathbf{T}(\mathbf{p}_x) - \mathbf{q}_y\right\|_2 < \tau_2 \right] \quad \in [0,1]\,, \\
FMR &= \frac{1}{M} \sum_{i=1}^{M} \left[IR > \tau_3 \right] \quad \in [0,1]\,,
\end{aligned} \tag{6}
$$

in which an inlier is defined as the distance between the two points is lower than a certain threshold τ_2 under the ground-truth transformation $\mathbf{T}$.

Relative Rotation Error (RRE) measures the geodesic distance in degrees between the estimated $\tilde{\mathbf{R}}$ and ground truth rotation $\mathbf{R}$ matrices:

$$
RRE = \arccos\left(\frac{trace(\mathbf{R}^T\tilde{\mathbf{R}}) - 1}{2}\right). \tag{7}
$$

Relative Translation Error (RTE) measures the Euclidean distance between the estimated $\tilde{\mathbf{t}}$ and ground truth translation $\mathbf{t}$ vectors:

$$
RTE = \left\|\mathbf{t} - \tilde{\mathbf{t}}\right\|. \tag{8}
$$

Modified Chamfer Distance measures the average of the pair-wise nearest distance between two point sets P and Q:

$$
\begin{aligned}
CD = &\frac{1}{|P|} \sum_{p\in P} \min_{q\in Q} \left\|\tilde{\mathbf{T}}(\mathbf{p}) - \mathbf{q}\right\|_2^2 + \\
&\frac{1}{|Q|} \sum_{q\in Q} \min_{p\in P} \left\|\mathbf{q} - \tilde{\mathbf{T}}(\mathbf{p})\right\|_2^2
\end{aligned} \tag{9}
$$

B.4. Point Cloud Mosaicking

Having the ground truth point cloud P and reconstructed point cloud P^*. The **Reconstruction Accuracy (Acc)** and **Reconstruction Completeness (Comp)** are defined as:

$$
\begin{aligned}
Acc &= \frac{1}{n} \sum_{p\in P}^{n} \min_{p^*\in P^*} (\|p - p^*\|) \\
Comp &= \frac{1}{n} \sum_{p^*\in P^*}^{n} \min_{p\in P} (\|p - p^*\|)
\end{aligned} \tag{10}
$$

And the **Reconstruction Precision (Prec.)** and **recall (Recall)** and the **F1-score** are defined as:

$$
\begin{aligned}
Prec. &= \frac{1}{n} \sum_{p\in P}^{n} \min_{p^*\in P^*} \left[\|p - p^*\| < 0.05\right] \quad \in [0,1]\,, \\
Recall &= \frac{1}{n} \sum_{p^*\in P^*}^{n} \min_{p\in P} \left[\|p - p^*\| < 0.05\right] \quad \in [0,1]\,, \\
F1 &= 2\frac{Prec. \times Recall}{Prec. + Recall} \quad \in [0,1]\,.
\end{aligned} \tag{11}
$$

C. Evaluation on Scene Graph Alignment with Controlled Semantic Noise and with Predicted 3D Scene Graph

We also test the robustness of our network against controlled noise on scene graph node alignment. Following the same implementation of SGAligner [34], we evaluate our method with 5 different types of noises: (i) only relationships are removed; (ii) only object(node) are removed their corresponding attributes and any relationships that include them are also removed; (iii) both relationships and object nodes are removed; (iv) object instances assigned with the wrong semantic label); and (v) both relationships and objects are both assigned with wrong semantics. Results are given in Table 8. We also list the noise-free result here as a reference.

Noise Types	Mean RR	F1	Hits @		
			K=1	K=3	K=5
(i)	96.70	77.52	94.93	98.56	98.80
(ii)	97.81	78.41	96.02	99.69	99.94
(iii)	96.86	77.15	94.43	99.35	99.89
(iv)	85.18	69.71	77.99	90.69	94.75
(v)	85.14	69.05	77.81	90.57	95.02
noise-free	97.91	88.39	96.24	99.66	99.93

Table 8. **Evaluation on node matching with different variants of controlled semantic noise**.

Our method shows very strong robustness against missing relationships (edges) and missing instances (nodes). In (iv) and (v), wrong instance semantic information shows relatively strong impacts on the alignment performance compared to wrong relationships. For testing the use of predicted 3D scene graphs instead of ground truth graphs, we generated predicted 3D scene graphs using [41] and tested our network (only trained on the ground truth) on the alignment task. Since the authors of [34] did not publish their code or pre-trained model for using predicted 3D scene graph, we **cannot guarantee a fair comparison** with their results. Table 9 reproduces theirs as in [34] compared with **ours on our validation set**.

Methods	Mean RR	F1	Hits @		
			K=1	K=3	K=5
SGA [34]	88.2	-	83.3	91.8	95.1
B+P+K	95.9	86.0	93.1	98.6	99.4

Table 9. **Evaluation on node matching with predicted graph.**

D. Additional Ablation Study

D.1. Object-per-Object Registration with Ours

Same as SGAligner [34], we conduct object-per-object point cloud registration following with RANSAC using the scene graph alignment results of our own network. To further improve the robustness of the object-to-object registration, we propose two methods: (1) The dense scene graph alignment result **S** is first filtered with a confidence threshold s, only when the score of object pairs is higher than s will be considered in point cloud registration. If none of the object pairs has a score higher than s, all object pairs are taken for registration, and (2) only top-k-scored object pairs will be used in registration. We also give the registration results of using our network with overlap-to-overlap (O2O) and using SGAligner (S*.) with O2O as references in Table 10. Our network combined with OPO registration performs marginally worse than with O2O registration, while for SGAligner the situation is the converse.

Methods	CD	RRE	RTE	FMR	RR
$s=0$	0.0544	4.9849	12.31	99.37	96.00
$s=0.3$	0.0581	4.8246	12.74	99.37	95.74
$s=0.5$	0.0462	3.9634	9.74	99.26	96.39
$k=3$	0.0627	5.1250	13.61	99.37	95.95
$k=5$	0.0514	4.7141	11.76	99.37	96.27
$k=7$	0.0574	5.0628	12.97	99.37	95.90
O2O	**0.0083**	**0.6252**	**1.32**	**99.73**	**99.57**
S*. + O2O	0.0179	1.3428	2.67	99.26	98.95

Table 10. **Object-per-Object Point Cloud Registration with our method.** Methods with s represent filter object pairs with confidence scores lower than the threshold, while methods with k take only the top-k object pairs for registration.

D.2. Fusion with Different Levels of Point Feature

KPConv-FPN [38] provides multi-level point geometric features of a point cloud. In the original implementation of Geotransformer, there are three levels of geometric features: coarse-level $N_c \times 1024$, middle-level $N_m \times 512$ and fine-level $N_f \times 256$. Here we give a comparison of using different levels of geometric features for the P2SG fusion module in terms of 3D scene graph alignment in Table 11. As the result shows, P2SG fusion with fine-level geometric features performs the best among all listed variants.

Methods	Mean RR	F1	Hits @		
			K=1	K=3	K=5
Coarse	97.00	85.51	94.69	99.33	99.79
Middle	97.85	87.67	96.24	99.58	99.83
Fine	**98.58**	**89.39**	**97.49**	**99.68**	**99.90**

Table 11. **Evaluation on node matching with different levels of point geometric feature.**

D.3. Alignment with Augmented Transformation

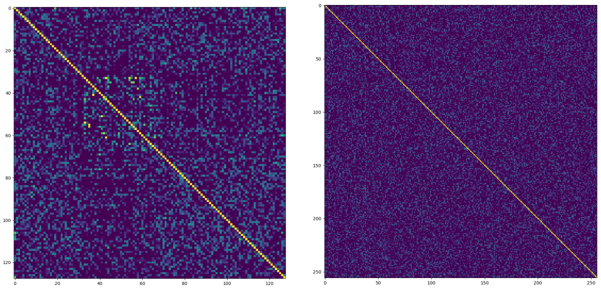
Here we provide the 3D scene graph alignment results with augmented T in Table 12 as the complementary of Figure 6.

Mtds.	Overlap (%)	Mean RR	F1	Hits @		
				K=1	K=3	K=5
SG-PGM (*ours*)	10-30	94.96	74.86	91.23	98.69	99.65
	30-60	97.91	87.95	96.33	99.54	99.87
	60-	99.15	95.21	98.48	99.83	99.93
	overall	**97.81**	**88.18**	**96.16**	**99.49**	**99.85**
SGA* [34]	10-30	79.93	60.46	64.64	86.54	93.50
	30-60	83.20	71.84	71.25	89.61	95.28
	60-	87.24	81.05	78.01	93.75	97.48
	overall	85.92	79.46	77.69	88.07	93.71

Table 12. **Evaluation of our proposed method on node matching per overlap range.** Even in low-overlap cases, our method still provides accurate alignment results with Hit@1 over 90%.

D.4. Analyse of AIS Module

Equation 2 gives the definition of the affinity matrix, in which the affinity of the embeddings from the scene graph and the point cloud is separately computed. In Figure 8, we provide a visualization of the learnable parameters W_s and W_p . As shown in the Figure, the multi-level scene graph embedding is more coupled crossing different feature channels, especially of the first-hop graph embedding, while the geometric feature is relatively more decoupled.



(a) W_s (b) W_p

Figure 8. **The learnable parameters W_s and W_p of the AIS Module.**

D.5. Additional comparison with GCNet on point cloud registration and overlap checking

We tested GCNet [56] on the registration task on our validation set in Table 13. We additionally combined our method with GCNet to mask out the feature points from unmatched objects before the Consistent Voting, which shows improvement compared to GCNet alone.

Methods	RRE	RTE	FMR	RR
GeoTr [31]	1.94	4.96	98.37	98.37
GeoTr + Ours	**1.57**	**3.51**	**99.47**	**98.72**
GCNet [56]	2.24	5.43	98.88	98.51
GCNet + Ours	1.96	4.91	99.09	98.72

Table 13. **Additional evaluation on point cloud registration.**

We also tested GCNet on the overlap checking task, using the average of the top 25% of predicted overlap score vector o and saliency score vector s. In Table 14, we report GCNet with $o_{25\%} \cdot s_{25\%} > 0.45$ as overlap, and the results of using the scene-level score k instead of Eq. 8 in our method. It shows a huge drop in Prec. because our partial graph matching module is only trained with overlapping samples.

Methods	Prec.	Recall	F1
SGA [34]	92.03	90.94	91.48
GCNet [56]	93.43	92.24	92.83
SG-PGM w/ $k > 0.45$	89.94	96.87	93.28
SG-PGM@3 (ours)	**95.41**	**95.01**	**95.21**

Table 14. **Overlap check for point cloud registration.**

E. Qualitative Results

Here we provide some qualitative results by combining our method and GeoTransformer [31] for point cloud registration in Figure 9 and for point cloud mosaicking in Figure 10.

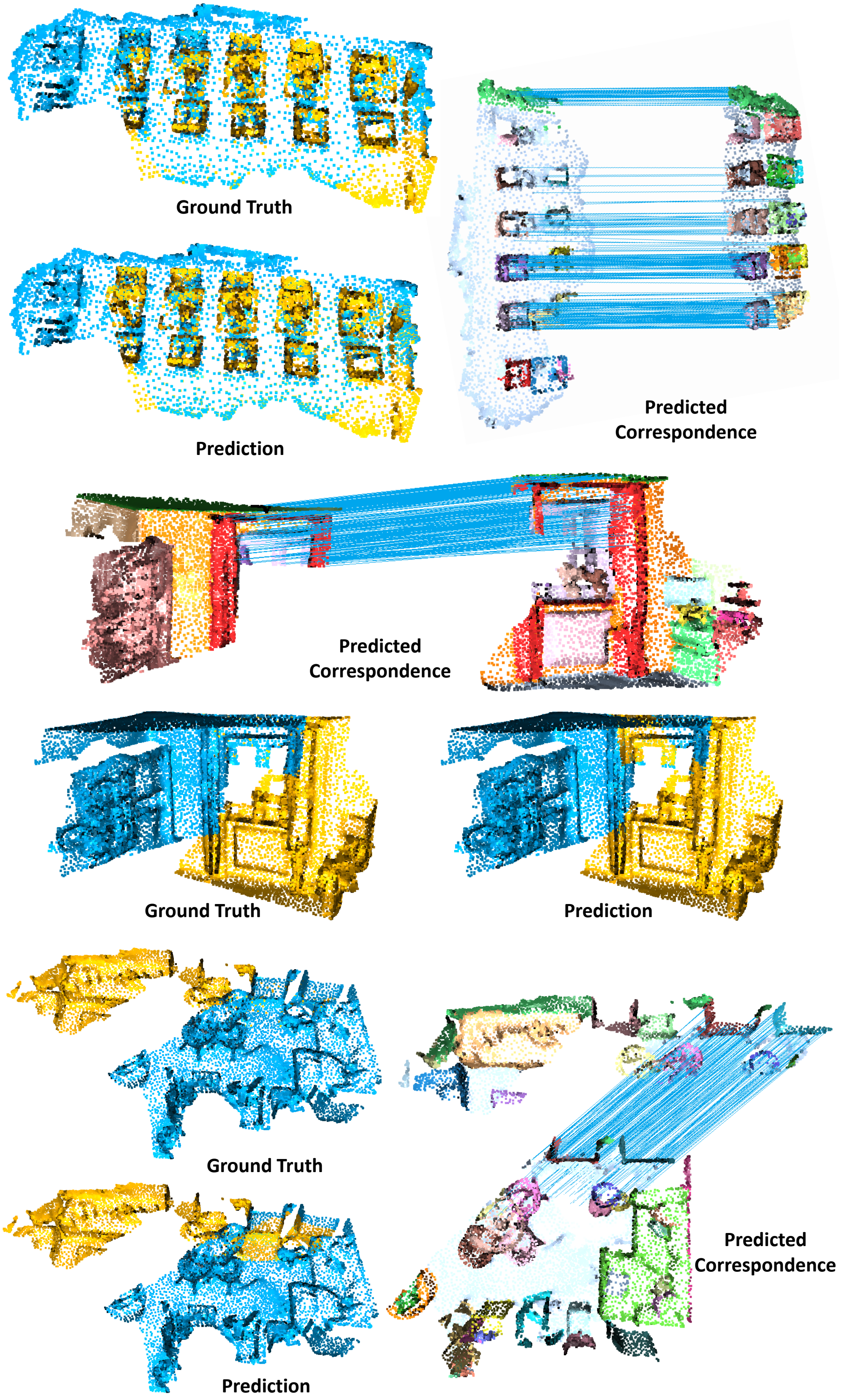


Figure 9. **Qualitative Results on Point Cloud Registration** of our proposed method.

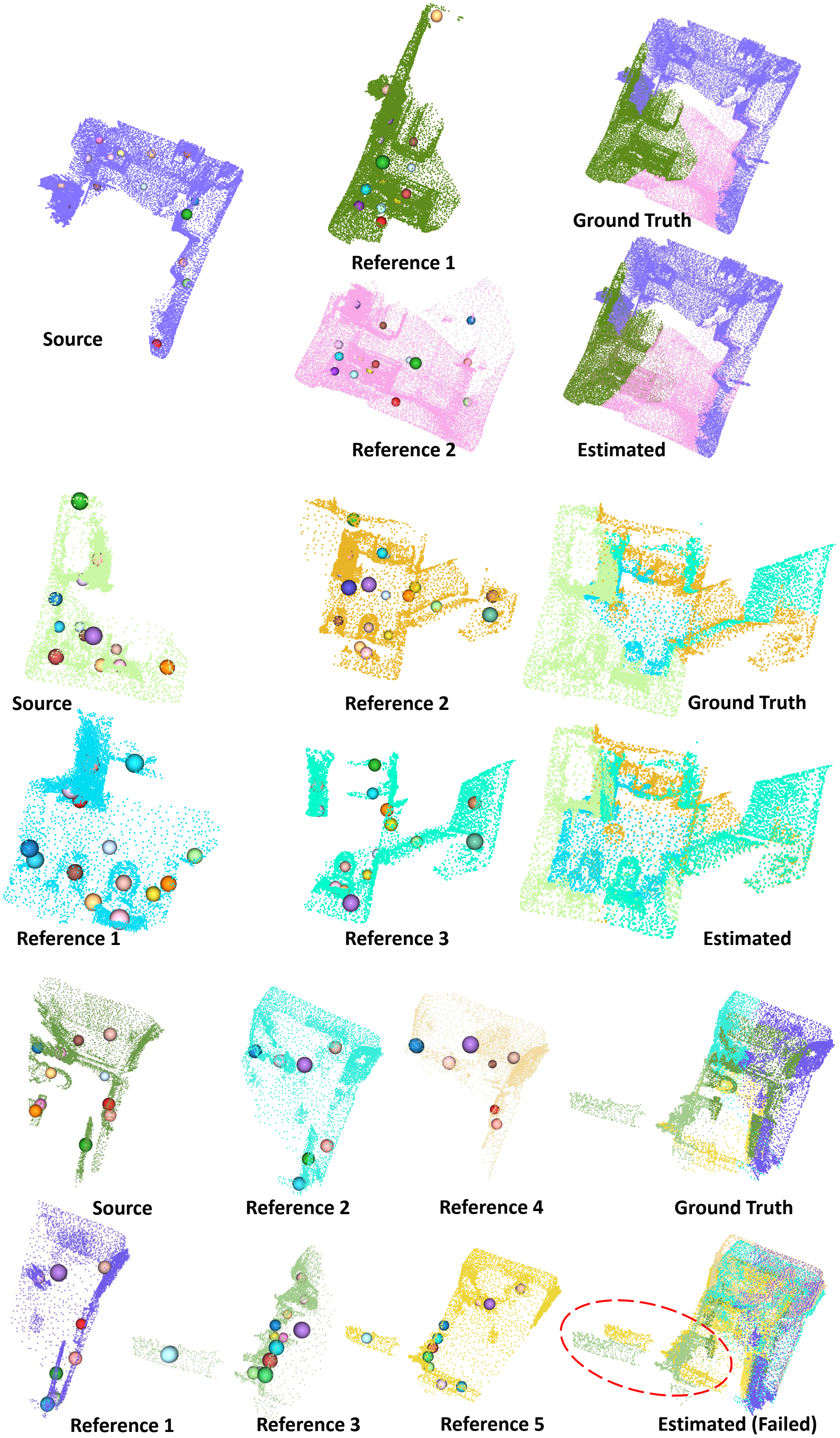


Figure 10. **Qualitative Results on Point Cloud Mosaicking** of our proposed method. Object nodes are visualized as 3D spheres.