CVPR Nashville JUNE 11-15, 2025

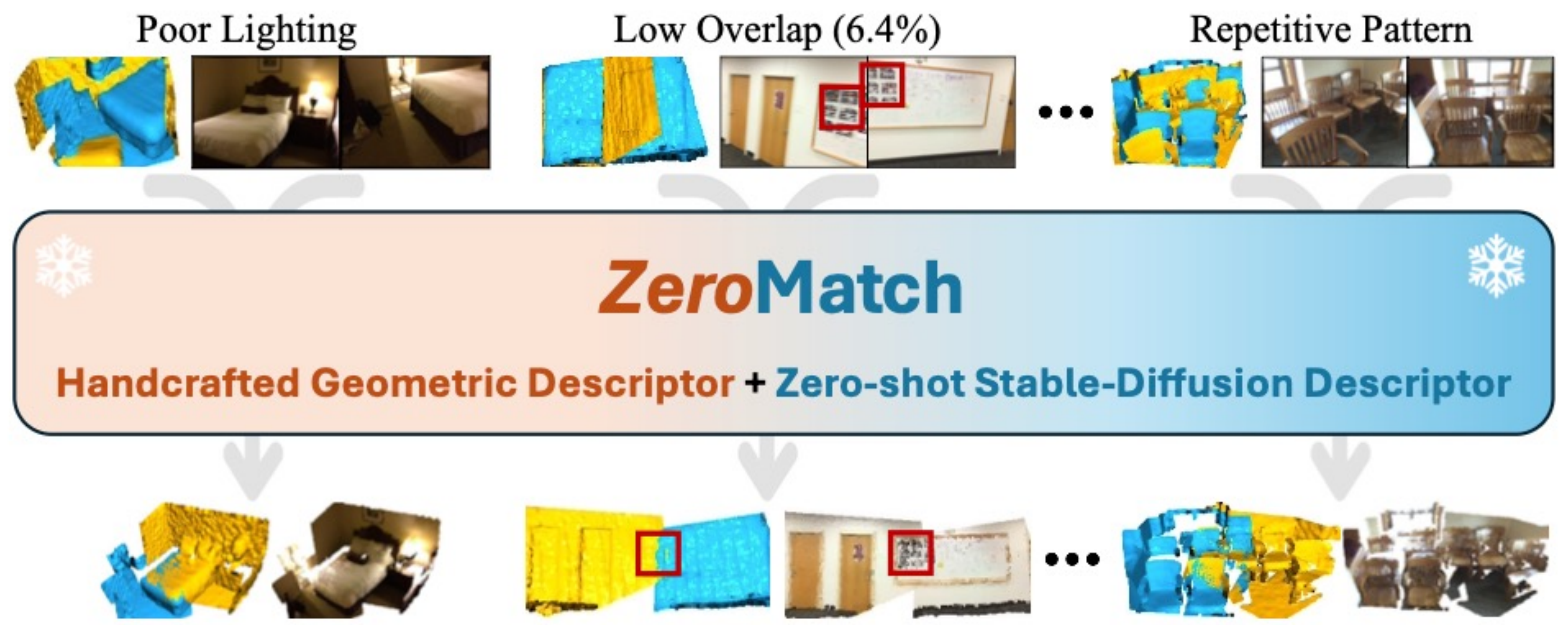
# Zero-Shot RGB-D Point Cloud Registration with Pre-trained Large Vision Model

Haobo Jiang[1], Jin Xie[4], Jian Yang[3], Liang Yu[2], Jianmin Zheng[1]

[1]Nanyang Technological University, [2]Alibaba Group, [3]Nankai University, [4]Nanjing University

## Background and Motivations

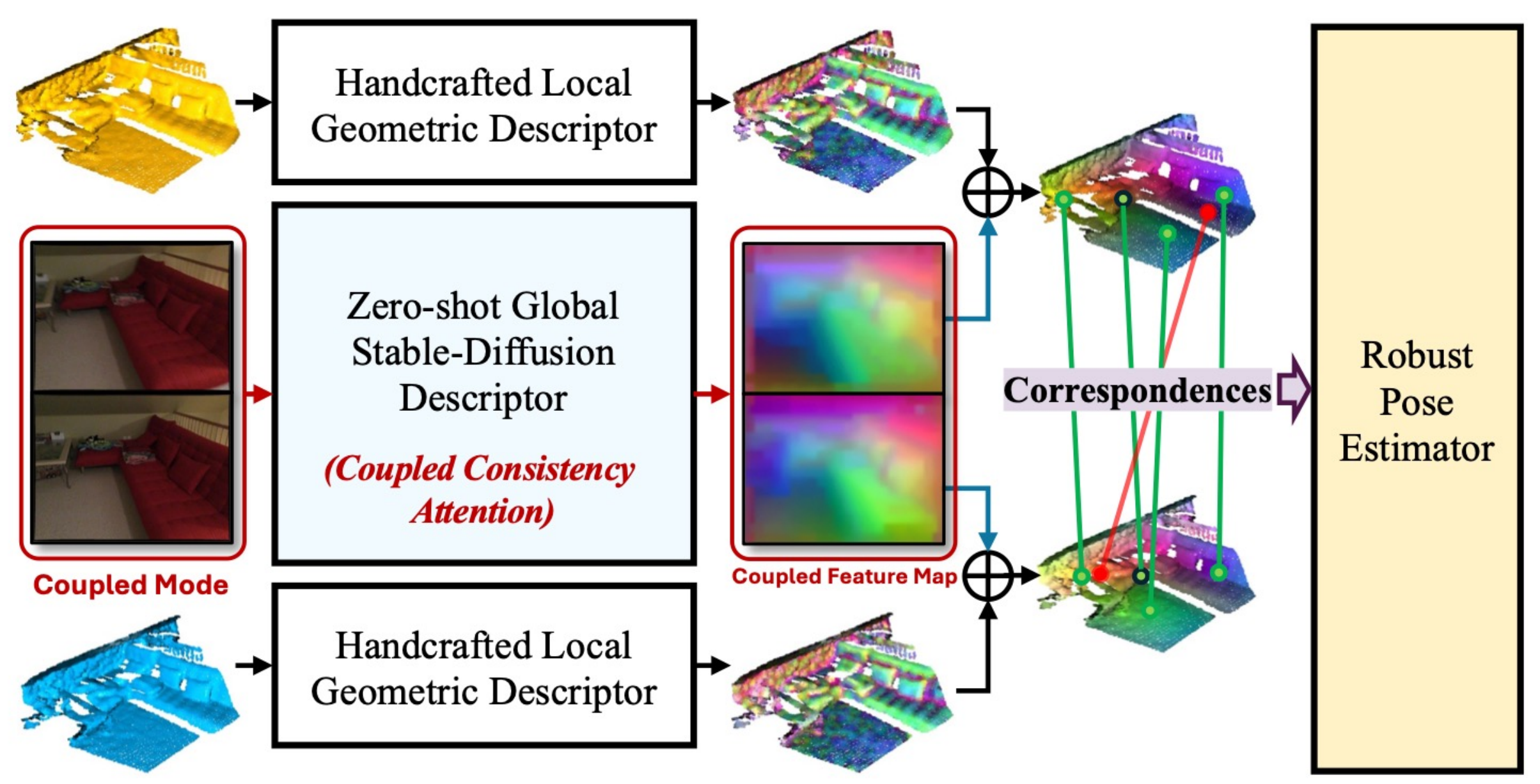
- ➢Traditional methods: Integrate color cues into ICP optimization; (Suffer from unreliable performances)
- ➢Deep methods: Deeply fuse semantic and geometric features; (Limited by training data, causing unstable generalization ability)
- ➢We propose **ZeroMatch,** a zero-shot RGB-D registration method driven by large vision model (i.e., Stable Diffusion);
- ➢**Motivation:** Leverage powerful zero-shot image representations from Stable Diffusion, achieved through extensive pre-training on large-scale data, to enhance point-cloud geometric descriptors.


Poor Lighting
Low Overlap (6.4%)
Repetitive Pattern
...
ZeroMatch
Handcrafted Geometric Descriptor + Zero-shot Stable-Diffusion Descriptor
...

## Contributions

- ➢We develop a novel zero-shot RGB-D 3D registration framework, **ZeroMatch**, leveraging the powerful zero-shot representations of Stable Diffusion to enhance handcrafted geometric descriptors for robust matching;
- ➢To enhance cross-view SD feature consistency, we propose a novel coupled-image input mode to replace the original single-image mode, enabling **inter-image and prompt-to-image consistency attentions** for robust cross-view feature alignment.

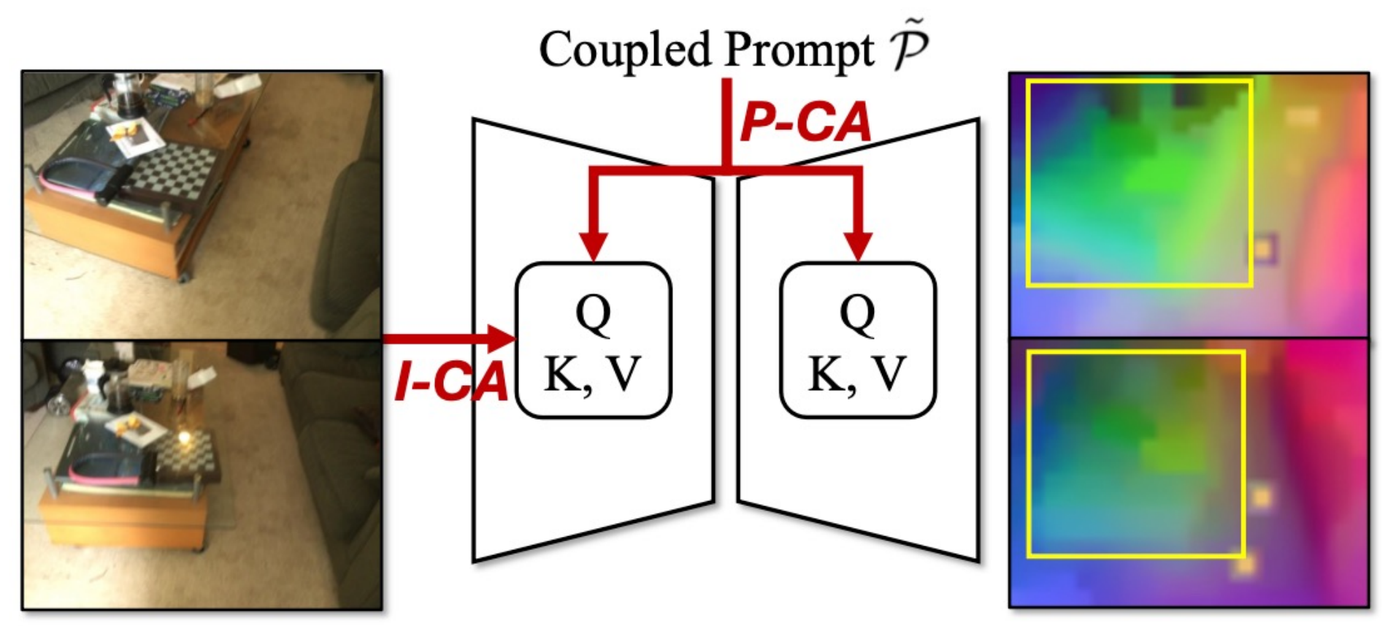
## Methodology


Handcrafted Local Geometric Descriptor
Zero-shot Global Stable-Diffusion Descriptor
(Coupled Consistency Attention)
Handcrafted Local Geometric Descriptor
Coupled Mode
Coupled Feature Map
Correspondences
Robust Pose Estimator

➢Our ZeroMatch consists of three core components:

- **Handcrafted local geometric descriptor** for characterizing local geometry of point clouds;
- **Zero-shot global Stable-Diffusion descriptor** for capturing global contextual information;
- **Local & global feature fusion and matching** for pose estimate.

***Coupled-Image Input Mode***



- ➢ **Image-level consistency attention (*I-CA*)**
- ➢ **Prompt-to-image consistency attention (*P-CA*)**

## Experiments

Table 1. Comparison of the methods on rotation, translation, and Chamfer distance on **ScanNet** [8] benchmark dataset.

| | | **Rotation** (*deg*) | | | | | **Translation** (*cm*) | | | | | **Chamfer** (*mm*) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy ↑ | | | Error↓ | | Accuracy ↑ | | | Error↓ | | Accuracy ↑ | | | Error↓ | |
| Methods | Train Set | 5 | 10 | 45 | Mean | Med. | 5 | 10 | 25 | Mean | Med. | 1 | 5 | 10 | Mean | Med. |
| ICP [4] | - | 31.7 | 55.6 | 99.6 | 10.4 | 8.8 | 7.5 | 19.4 | 74.6 | 22.4 | 20.0 | 8.4 | 24.7 | 40.5 | 32.9 | 14.1 |
| SIFT [34] | - | 55.2 | 75.7 | 89.2 | 18.6 | 4.3 | 17.7 | 44.5 | 79.8 | 26.5 | 11.2 | 38.1 | 70.6 | 78.3 | 42.6 | 1.7 |
| SuperPoint [12] | - | 65.5 | 86.9 | 96.6 | 8.9 | 3.6 | 21.2 | 51.7 | 88.0 | 16.1 | 9.7 | 45.7 | 81.1 | 88.2 | 19.2 | 1.2 |
| FCGF [6] | 3DMatch | 70.2 | 87.7 | 96.2 | 9.5 | 3.3 | 27.5 | 58.3 | 82.9 | 23.6 | 8.3 | 52.0 | 78.0 | 83.7 | 24.4 | 0.9 |
| DGR [7] | 3DMatch | 81.1 | 89.3 | 94.8 | 9.4 | 1.8 | 54.5 | 76.2 | 88.7 | 18.7 | 4.5 | 70.5 | 85.5 | 89.0 | 13.7 | 0.4 |
| 3D MV Reg [19] | 3DMatch | 87.7 | 93.2 | 97.0 | 6.0 | 1.2 | 69.0 | 83.1 | 91.8 | 11.7 | 2.9 | 78.9 | 89.2 | 91.8 | 10.2 | 0.2 |
| REGTR [52] | 3DMatch | 86.0 | 93.9 | 98.6 | 4.4 | 1.6 | 61.4 | 80.3 | 91.4 | 14.4 | 3.8 | 80.9 | 90.9 | 93.6 | 13.5 | 0.2 |
| GeoTransformer [40] | 3DMatch | 94.0 | 96.8 | 98.1 | 4.3 | 1.0 | 79.2 | 92.0 | 96.7 | 8.2 | 2.5 | 88.4 | 95.8 | 96.9 | 5.8 | **0.1** |
| PEAL [53] | 3DMatch | 94.4 | 96.8 | 98.4 | 3.9 | 0.9 | 80.5 | 92.8 | 97.0 | 7.3 | 2.4 | 89.1 | 96.0 | 97.1 | 6.0 | **0.1** |
| UR&R (RGB-D) [15] | 3DMatch | 87.6 | 93.1 | 98.3 | 9.8 | 3.1 | 69.2 | 84.0 | 93.8 | 9.5 | 2.8 | 79.7 | 91.3 | 94.0 | 9.3 | 0.2 |
| UR&R [15] | 3DMatch | 87.6 | 93.7 | 98.3 | 9.8 | 3.8 | 67.5 | 83.8 | 94.6 | 9.5 | 3.0 | 91.7 | 94.6 | 94.6 | 6.5 | 0.2 |
| BYOC [14] | 3DMatch | 66.5 | 85.2 | 97.8 | 7.4 | 3.3 | 30.7 | 57.6 | 88.9 | 16.0 | 8.2 | 54.1 | 82.8 | 89.5 | 9.5 | 0.9 |
| LLT [50] | 3DMatch | 93.4 | 96.5 | 98.2 | 2.5 | 0.8 | 76.9 | 90.2 | 96.7 | 5.5 | 2.2 | 86.4 | 95.1 | 95.8 | 4.6 | **0.1** |
| PointMBF [27] | 3DMatch | 94.6 | 97.0 | 98.7 | 3.0 | 0.8 | 81.0 | 92.0 | 97.1 | 6.2 | 2.1 | 91.3 | 96.6 | 97.4 | 4.9 | **0.1** |
| NeRF-UR [54] | 3DMatch | 97.2 | 99.0 | 99.7 | 1.6 | 0.9 | 84.2 | 95.8 | 98.7 | 3.9 | 2.2 | 93.2 | 98.3 | 98.8 | 2.7 | **0.1** |
| UR&R [15] | ScanNet | 92.7 | 95.8 | 98.5 | 3.4 | 0.8 | 77.2 | 89.6 | 96.1 | 7.3 | 2.3 | 86.0 | 94.6 | 96.1 | 5.9 | **0.1** |
| UR&R (RGB-D) [15] | ScanNet | 94.1 | 97.0 | 99.1 | 2.6 | 0.8 | 78.4 | 91.1 | 97.3 | 5.9 | 2.3 | 87.3 | 95.6 | 97.2 | 5.0 | **0.1** |
| BYOC [14] | ScanNet | 86.5 | 95.2 | 99.1 | 3.8 | 1.7 | 56.4 | 80.6 | 96.3 | 8.7 | 4.3 | 78.1 | 93.9 | 96.4 | 5.6 | 0.3 |
| LLT [50] | ScanNet | 95.5 | 97.6 | 99.1 | 2.5 | 0.8 | 80.4 | 92.2 | 97.6 | 5.5 | 2.2 | 88.9 | 96.4 | 97.6 | 4.6 | **0.1** |
| PointMBF [27] | ScanNet | 96.0 | 97.6 | 98.9 | 2.5 | **0.7** | 83.9 | 93.8 | 97.7 | 5.6 | 1.9 | 92.8 | 97.3 | 97.9 | 4.7 | **0.1** |
| NeRF-UR [54] | ScanNet | 97.8 | 99.2 | 99.8 | 1.4 | 0.8 | 86.9 | 96.3 | **98.9** | 3.6 | 2.0 | 94.3 | 98.5 | 99.0 | **2.6** | **0.1** |
| **ZeroMatch** | - | **98.9** | **99.6** | **99.9** | **1.1** | **0.7** | **89.9** | **96.4** | 98.8 | **3.1** | **1.6** | **95.6** | **98.6** | **99.1** | 3.0 | **0.1** |

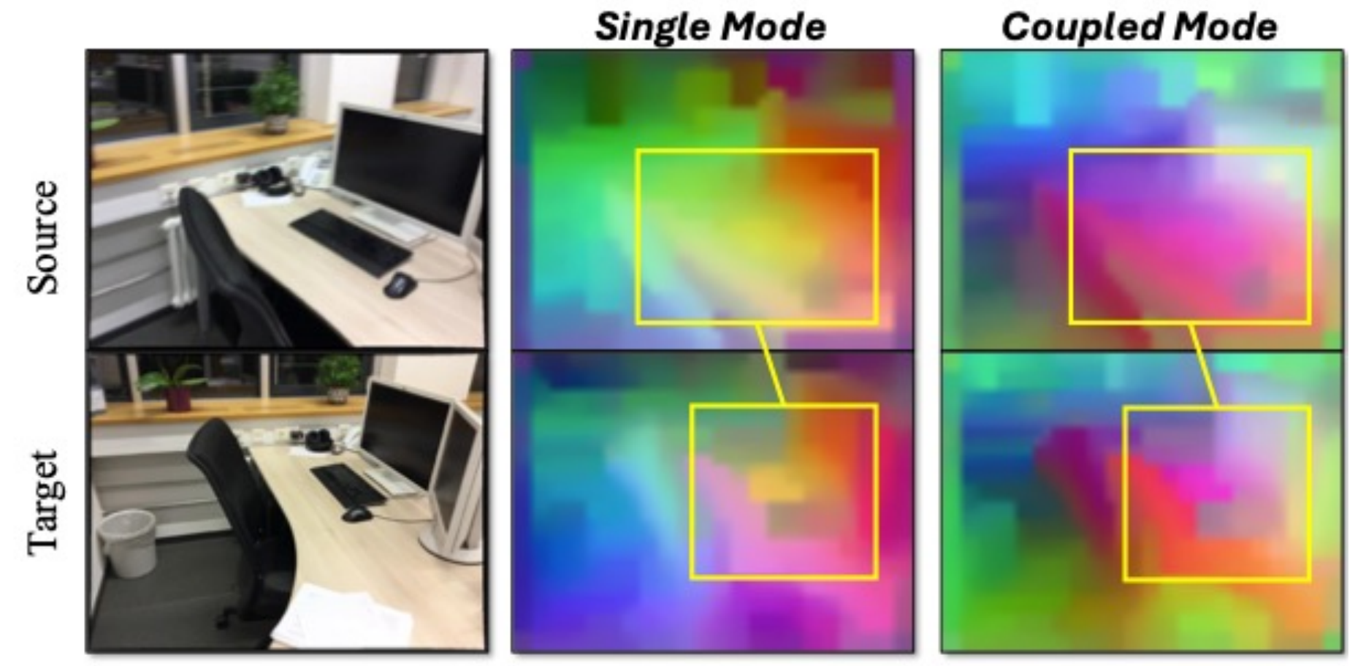


Table 4. Ablation studies about coupled-image input mode and coupled consistency attention on ScanLoNet dataset [8].

| | Accuracy ↑ | | | Error↓ | |
|---|---|---|---|---|---|
| **Chamfer** (*mm*) | 1 | 5 | 10 | Mean | Med. |
| ZeroMatch w/ Single | 66.5 | 81.6 | 84.6 | 60.1 | **0.4** |
| ZeroMatch w/ Coupled* | **67.5** | **82.0** | **84.7** | **58.6** | **0.4** |
| ZeroMatch w/o P-CA | 66.1 | 81.3 | 84.2 | 60.2 | 0.5 |
| ZeroMatch w/o I-CA | 67.4 | 81.5 | 84.3 | 61.3 | **0.4** |
| ZeroMatch* | **67.5** | **82.0** | **84.7** | **58.6** | **0.4** |
| ZeroMatch ($\alpha = 0.6$) | 67.1 | 81.2 | 84.0 | 59.5 | **0.4** |
| ZeroMatch ($\alpha = 0.7$) | 67.1 | 81.4 | 84.2 | 60.2 | **0.4** |
| ZeroMatch ($\alpha = 0.8$)* | **67.5** | **82.0** | 84.7 | 58.6 | **0.4** |
| ZeroMatch ($\alpha = 0.9$) | 67.3 | 81.8 | **84.8** | **56.6** | **0.4** |
| ZeroMatch ($\alpha = 1.0$) | 67.4 | 81.5 | 84.3 | 61.3 | **0.4** |

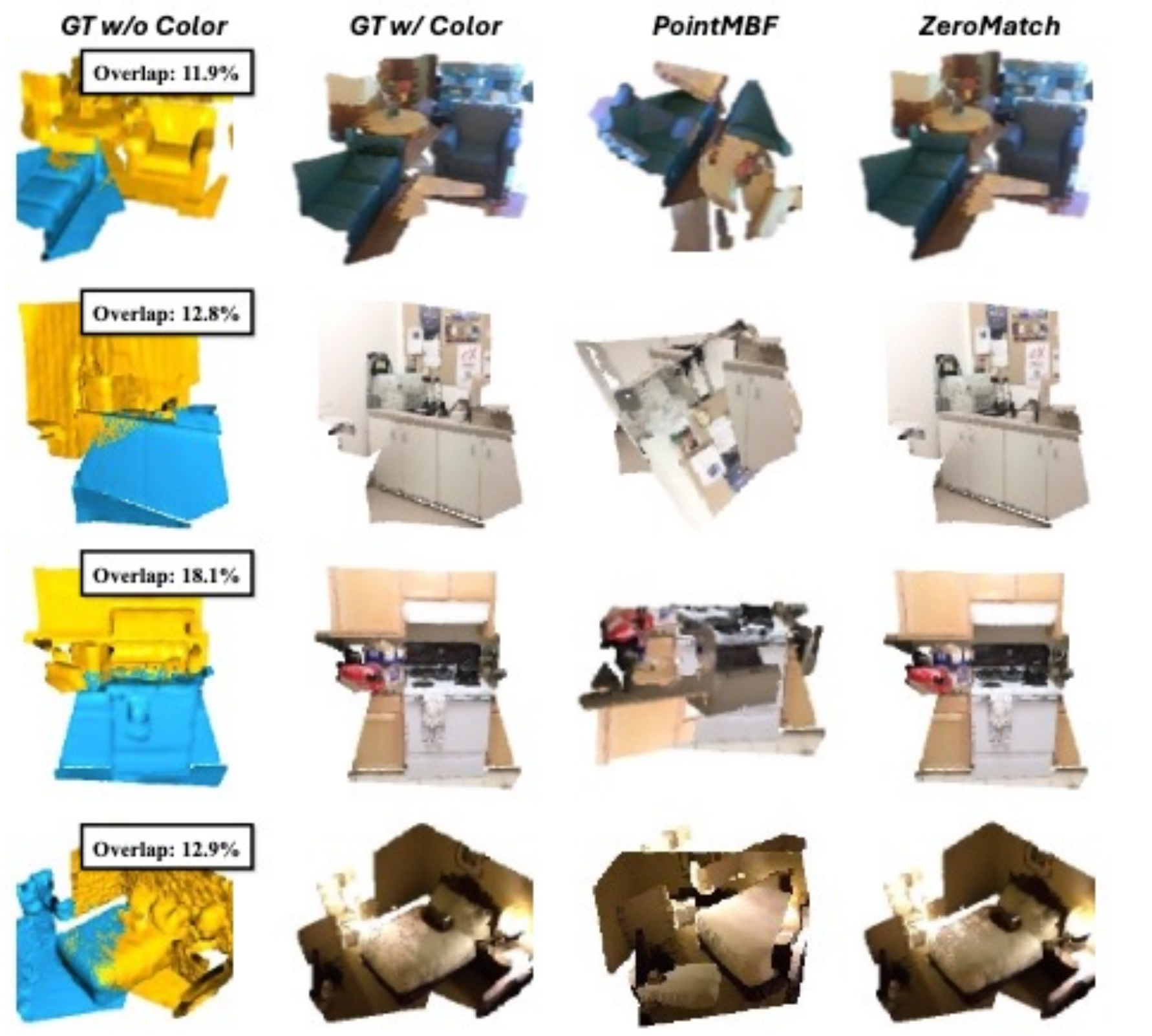


Figure 7. Qualitative comparisons on ScanLoNet dataset [8].

Table 5. Ablation studies about feature fusion on ScanLoNet [8].

| | Accuracy ↑ | | | Error↓ | |
|---|---|---|---|---|---|
| **Chamfer** (*mm*) | 1 | 5 | 10 | Mean | Med. |
| ZeroMatch w/ Geo | 60.1 | 72.2 | 75.1 | 121.0 | 0.5 |
| ZeroMatch w/ SD | 63.0 | 79.6 | 82.7 | 67.1 | 0.5 |
| ZeroMatch w/ Geo + SD* | **67.5** | **82.0** | **84.7** | **58.6** | **0.4** |