Supplementary Material for "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

{haobo.jiang, ASJMZheng}@ntu.edu.sg, liangyu.yl@alibaba-inc.com, csjyang@nankai.edu.cn, csjxie@nju.edu.cn

1. Qualitative Registration Comparisons

We provide additional qualitative comparisons with the state-of-the-art (SOTA) deep RGB-D registration method, PointMBF [2], on the challenging ScanLoNet [1] benchmark dataset, as shown in Fig. 1. In scenarios with low overlap and repetitive patterns, our method, ZeroMatch, demonstrates significantly greater robustness and precision compared to PointMBF. This improvement is primarily attributed to the effective local and global point cloud representations from the local geometric descriptors and the global Stable-Diffusion descriptors.

2. Stable-Diffusion Feature Comparisons

Furthermore, we present additional comparisons of Stable-Diffusion features using t-SNE visualizations for the single-image input mode and the coupled-image input mode, as shown in Fig. 2. It can be observed that our coupled mode exhibits superior cross-view feature consistency, where corresponding regions are represented with consistent colors, compared to the single mode. This advantage stems from the coupled mode's ability to introduce both image-level coupled consistency attention and prompt-to-image coupled consistency attention, enabling more effective cross-view feature interaction and alignment.

3. Failure Cases

Finally, we present some failure cases on the ScanLoNet dataset, as shown in Fig. 3. These cases demonstrate that when dealing with point cloud pairs with extremely low overlap, particularly when the overlap regions lack distinguishable geometric or texture features, ZeroMatch fails in registration. Additionally, since the ScanLoNet dataset generates view pairs with a large 50-frame interval, many view pairs have no overlapping regions, inevitably leading to registration failures for our method.

Limitation. The primary limitation of our method is the inference speed. Our average runtime per case on ScanLoNet is 1.48s (SD: 0.66s + Geo: 0.46s + Others: 0.36s), which, although offering significant performance advantages, is slower than UR&R (0.43s), LLT (0.47s), and PointMBF (0.34s). Future work will aim to accelerate Stable Diffusion feature extraction using model compression techniques, such as knowledge distillation. Please refer to supplementary material for additional failure cases.

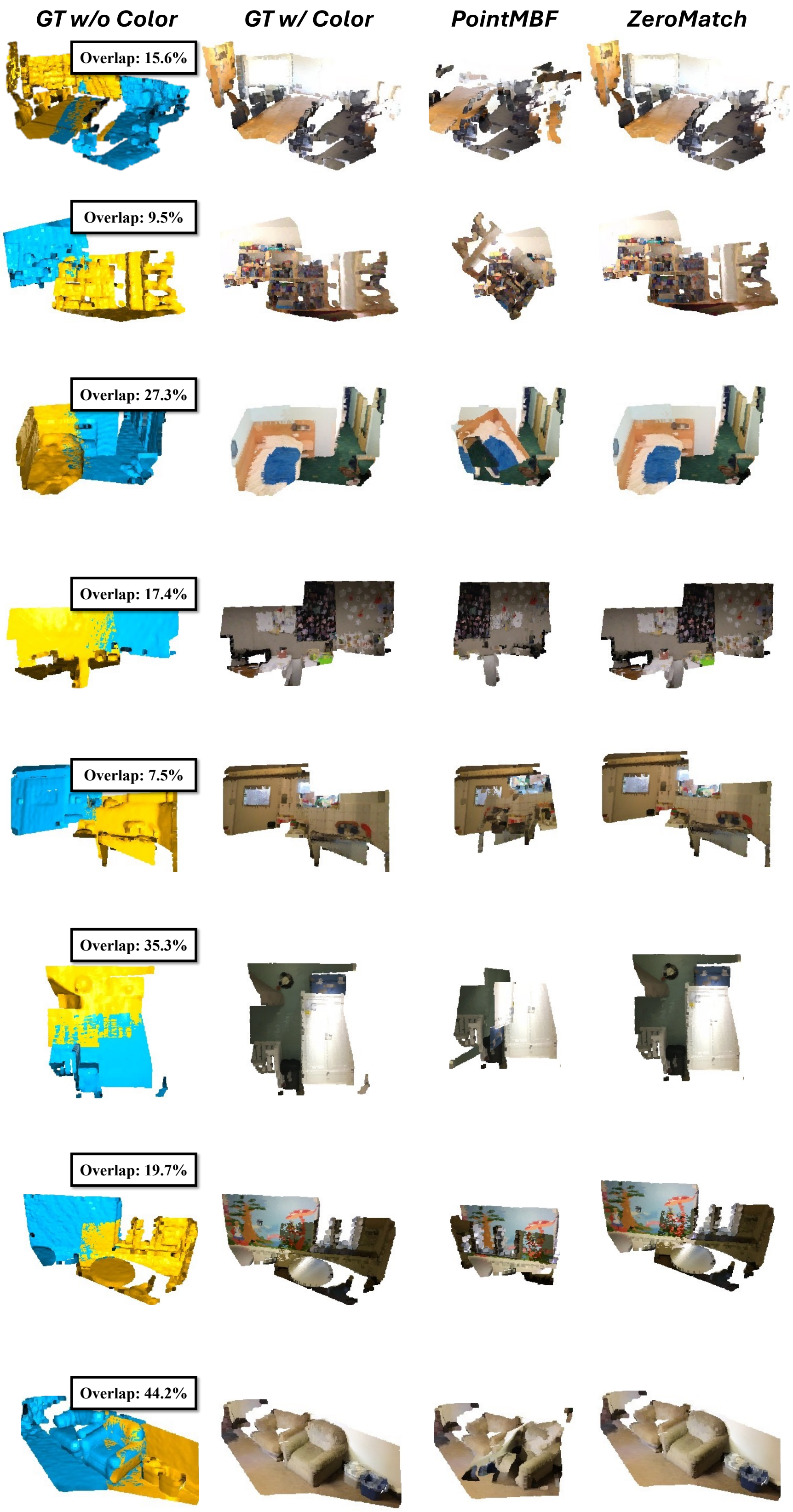

Figure 1. Qualitative comparisons on ScanLoNet dataset [1].

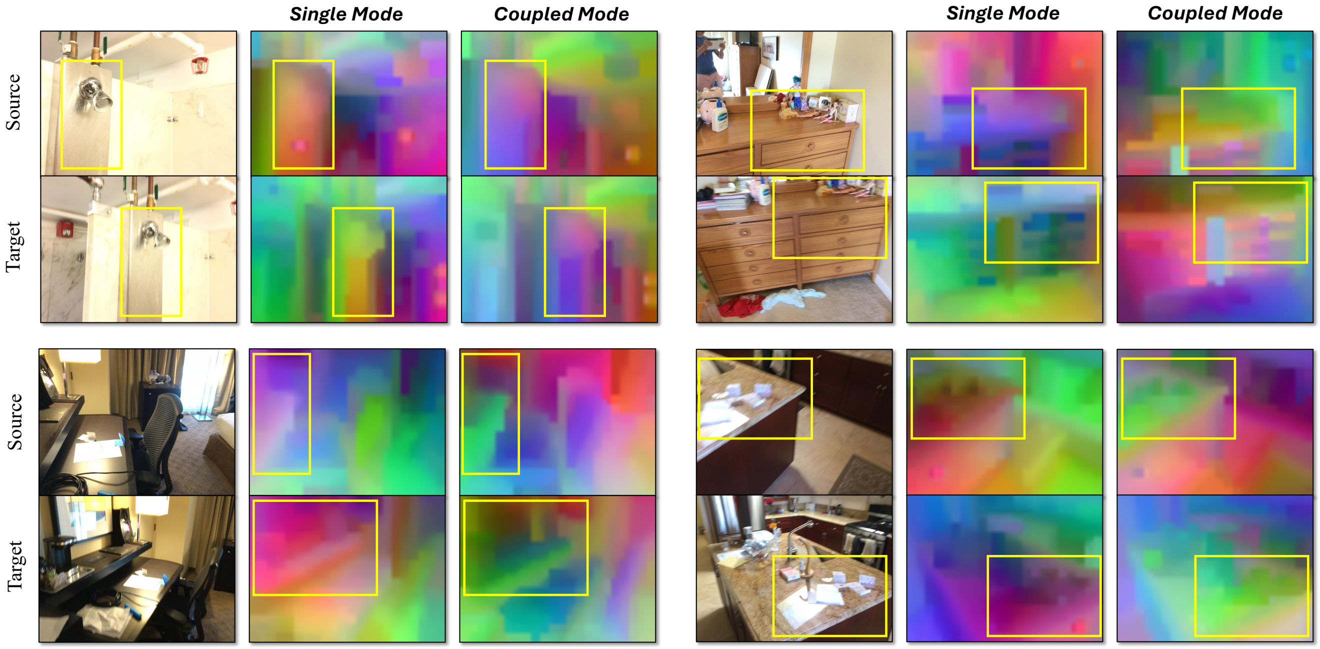

Figure 2. The visualized Stable-Diffusion features via t-SNE under the single-image and coupled-image input modes.

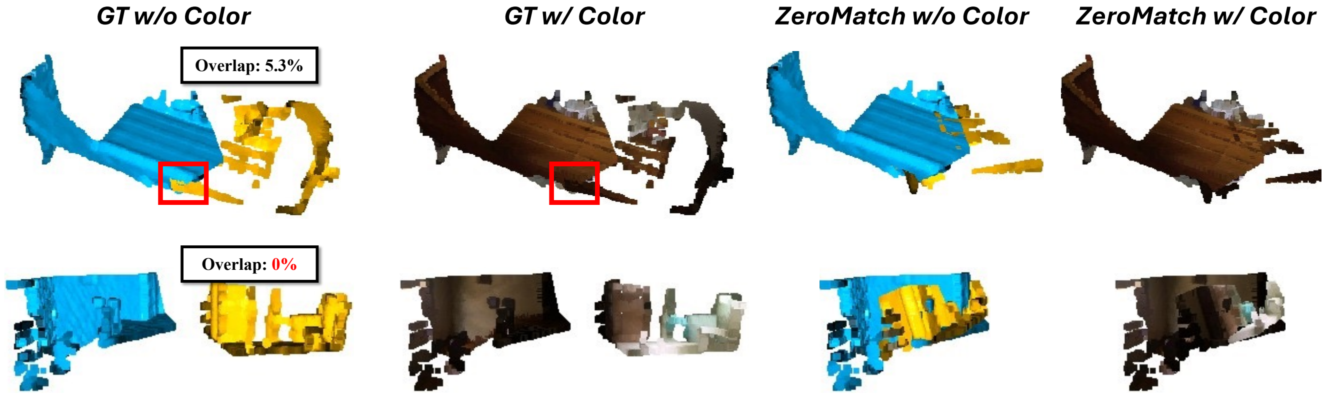

Figure 3. Failure cases in point cloud pairs with extremely low or no overlap.

References

[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 1, 2

[2] ICCV. Pointmbf: A multi-scale bidirectional fusion network for unsupervised rgb-d point cloud registration. 2023. 1