SurfelReg: Surfel-based 3D registration with Equivariant Features

Xueyang Kang & Hang Zhao & Zhaoliang Luan & Patrick Vandewalle & Kourosh Khoshelham

CVPR JUNE 17-21, 2024 SEATTLE, WA

THE UNIVERSITY OF MELBOURNE
KU LEUVEN

Abstract

Point cloud registration is a crucial challenge in the field of 3D reconstruction to ensure 3D alignment consistency. Despite the existence of various point cloud-based registration methods, both non-learning and learning-based, the utilization of surfels as geometry representation primitive, including both position and orientation, along with the associated uncertainty remains insufficiently explored. The point cloud based methods ignore the data uncertainty, and learn both rotation and translation implicitly from the point coordinates, leading to a reliance on dense training point clouds. To address these issues, we propose a novel surfel-based deep pose regression approach. Our method initializes surfels from depth map based on the specific depth camera projection model. Subsequently, the model learns both position and rotation representations through the $\mathbf{SE(3)}$-equivariant convolutional kernel for the relative transformation between paired frames. The model primarily consists of an equivariant convolutional encoder, a cross-attention mechanism for similarity computation, a fully-connected layer based decoder, and a non-linear Huber loss.

Surfel Init

Figure 1: In surfel init stage, the input is the depth map. Initially, we calculate the gradient of the depth map using the Sobel operator to derive the normal map, as illustrated in the middle. Subsequently, the downsampled depth map is transformed into a point cloud utilizing the camera projection. The point position corresponds to the center of a disk, the normal determines the orientation of the disk, and the disk radius represents the uncertainty.

The disk radius (standard deviation) is defined as below,

$$\epsilon_i = C\frac{e^{-\hat{\rho}}}{1+e^{-\tan(\theta)}}, \tag{1}$$

$$\theta = \arccos\left(\frac{\vec{r}\cdot\vec{o}}{\|\vec{r}\|\|\vec{o}\|}\right). \tag{2}$$

$$\hat{\rho} = \min\left(\max\left(\rho, \rho_{min}\right), \rho_{max}\right). \tag{3}$$

Model Structure

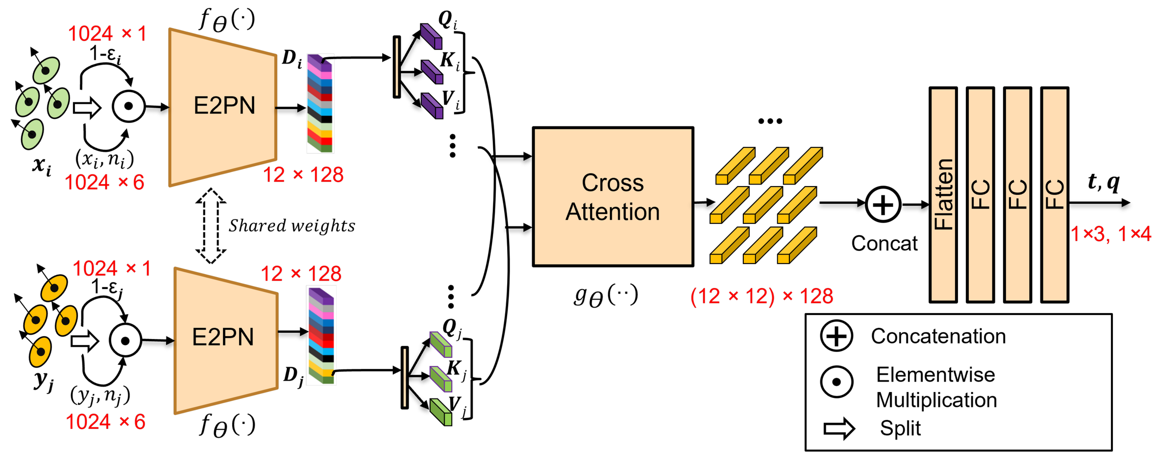

Figure 2: The network structure includes a shared encoder for surfels (split into 6 plus the uncertainty radius) from both the source and target frames in the $\mathbf{SE(3)}$ space. This encoder maps the 1024 surfels with 6 dimensions (position and normal) after weighting by confidence value $(1-\epsilon(\cdot))$ into 128-dimensional features across 12 channels, and then each descriptor along the channel dimension undergoes linear embedding to produce triplet token embeddings, denoted as $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$. The cross-attention $g_\theta(\cdot)$ is applied to feature descriptors from the source and the target frame. The resulting latent tokens are then combined to create a 2D feature map with dimensions $(12\times12)\times128$, as specified within the brackets following the cross-attention module. This feature map is later flattened into 1D and passed through Fully-Connected (FC) layers. These layers map the features into the relative position and the relative rotation in quaternion.

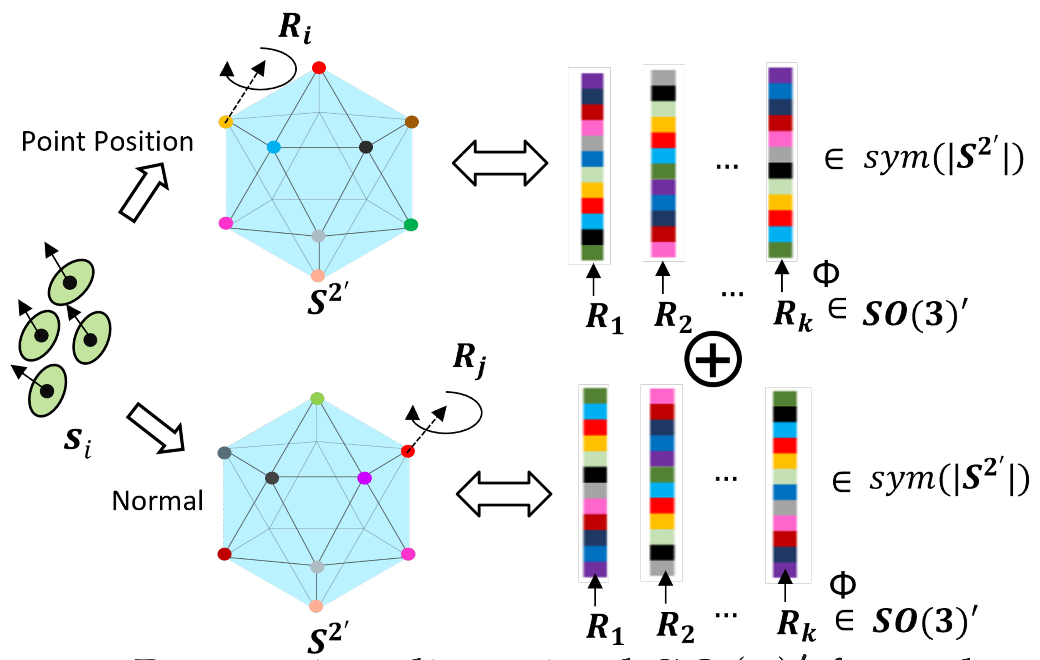

Figure 3: Recovering discretized $\mathbf{SO(3)}'$ from the quotient feature $S^{2'}$ by permutation order of points on Platonic solids of 12 corners.

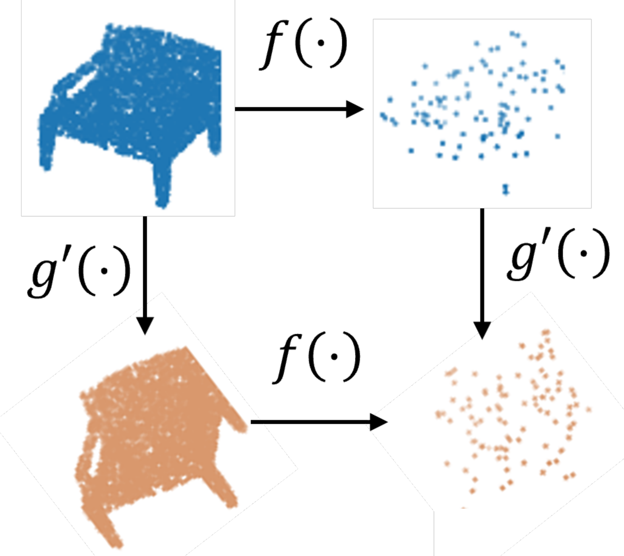

Figure 4: $f(\cdot)$, $g(\cdot)$ indicate network model and group rotation respectively.

Equivariant features of rotations and positions are learnt separately in two branches.

Experiment Results

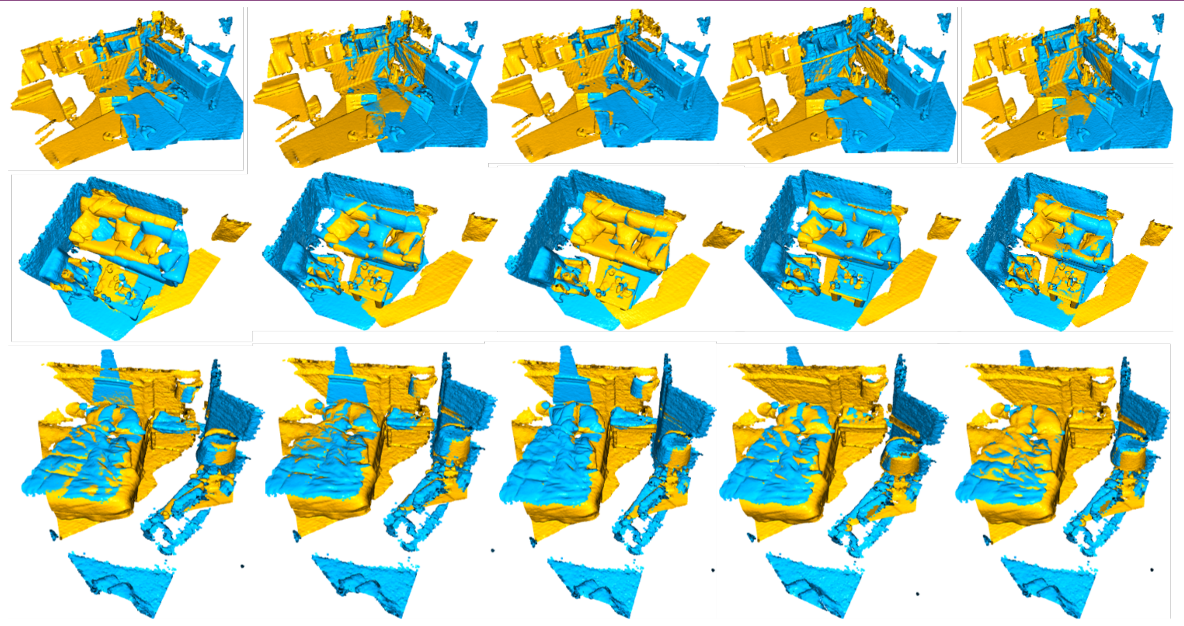

Figure 5: Comparisons on 3DMatch datasets. Yellow indicated the scan of source frame, while blue depicts the scan of target frame. The top three models with good performance are presented visually.

Method	ARKitScenes [4]				3DMatch [50]			
	RE(°)↓	TE(cm)↓	RR(%)↑	F1(%)↑	RE(°)↓	TE(cm)↓	RR(%)↑	F1(%)↑
FCGF [9]	2.17	7.69	90.47	89.71	2.31	7.06	91.72	89.14
DGR [10] ↑	1.74	6.13	93.95	90.86	2.40	7.48	91.30	89.76
D3Feat [2]	2.29	6.98	90.16	88.72.	2.57	8.16	89.79	87.40
SpinNet [1]	1.49	5.83	94.81	91.37	1.93	6.24	93.74	92.07
PointDSC [3]	1.62	6.08	94.54	91.18	2.06	6.55	93.28	89.35
RoReg [43] ↓	1.78	6.17	93.86	91.02	1.84	6.28	93.70	91.60
Ours	**1.37**	**5.72**	**95.08**	**93.32**	**1.57**	**6.09**	**94.05**	**92.63**

Method	RE(°)↓	TE(cm)↓	RR(%)↑
1. Surfel w/o uncertainty scaling	1.64	6.75	91.73
2. Point cloud + vanilla E2PN	2.24	7.16	89.85
3. W/o attention module	3.62	8.94	85.36
4. $\mathcal{L}_1$ loss	1.96	6.95	87.08
5. $\mathcal{L}_2$ loss	2.49	7.51	84.85
6. Binary cross entropy loss	1.84	6.69	90.17
7. Full surfel model	**1.57**	**6.09**	**94.05**

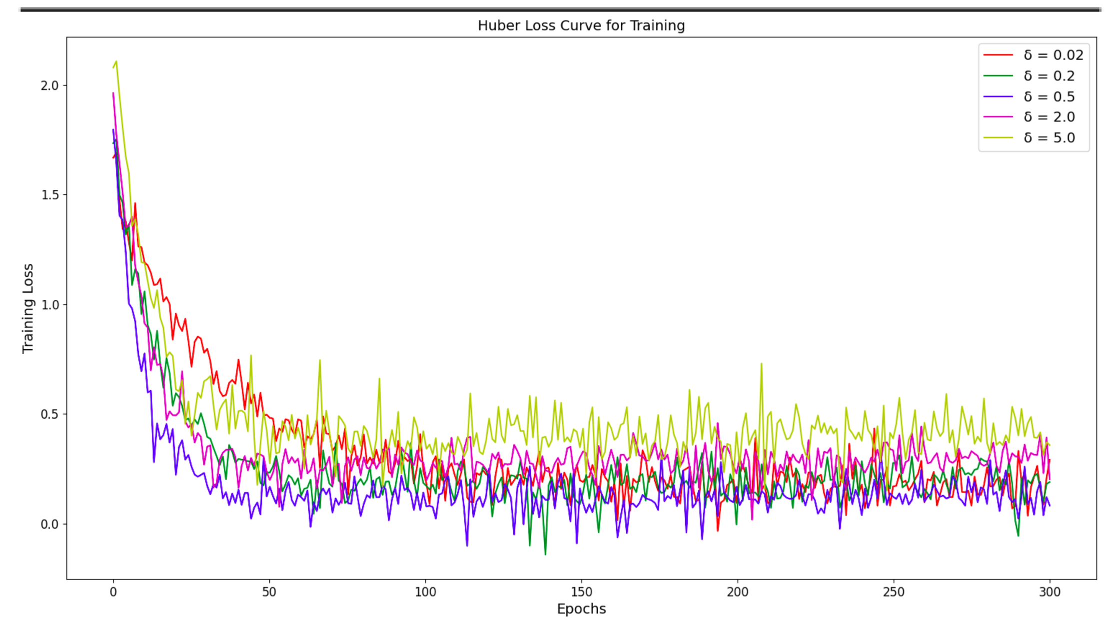

Figure 6: Huber loss learning curve under different thresholds.

Conclusion

In summary, the main contribution of this work can be briefly outlined as follows:

- Surfel initialized from the depth map is devised to generate the input for the pose regression model.
- A specialized deep learning model based on surfels is implemented to learn equivariant features using rotation-equivariant convolution.
- The differentiable Huber loss function is employed to explicitly leveraging the soft correspondence supervision of point pair candidates.

We present a complete surfel-based network model in conjunction with a surfel initialization method. Additionally, the camera view-based uncertainty initialization approach is devised. Our model consists of a shared E2PN encoder to learn equivariant features from the surfel of the source and target frames, a cross-attention module to establish feature correspondences, and the MLP-based decoder. Through the extensive comparison experiments and the ablation study on real 3D indoor scan datasets, we exhibit the model's robustness, good accuracy performance compared to the state-of-the-art models. However, despite the good performance on our selected scenes, the model may still have some limitations when the sampling points' density varies a lot spatially, different from the sampling from the uniform density distribution in 3D space. Lastly, the 2D gaussian surfel primitive can be explored further in the future for many other 3D tasks, like 3D mapping or 3D reconstruction.

Contact Information

Processing Speech & Images(PSI), Department of Electrical Engineering(ESAT) KU Leuven

Faculty of Engineering and Information Technology, the University of Melbourne.

Email alex.kang@kuleuven.be
xueyangk@student.unimelb.edu.au