USIP: Unsupervised Stable Interest Point Detection from 3D Point Clouds - Supplementary Materials

Jiaxin Li* Gim Hee Lee
Department of Computer Science, National University of Singapore

A. Overview

We provide more details on the algorithms and experiments described in the main paper. Sec. B presents more examples of the network degeneracy. Sec. C evaluates the effect of point-to-point loss $\mathcal{L}_p$ on the keypoint repeatability. Sec. D illustrates the details of our feature descriptor design. Sec. E gives more experiments on point cloud registration tasks. Sec. F presents visualizations of our USIP keypoints in various datasets.

B. More Examples on Degeneracy

As analyzed in Sec. 5, our FPN degenerates when the receptive field becomes sufficiently large, *i.e.*, it has gained sufficient global semantic information. The receptive field of the FPN is controlled by two parameters: number of keypoint proposals M and number of neighbors K in the KNN feature aggregation. More specifically, the receptive field size is proportional to K and inversely proportional to M. In this section, we visualize the network degeneracy by gradually enlarging the receptive field. Fig. 6 shows the degeneracies when $M = 64$ and $K = \{9, 24, 32, 40, 48, 64\}$. Fig. 7 shows the degeneracies when $K = 9$ and $M = \{64, 24, 20, 16, 12, 9\}$.

C. Effect of λ in Point-to-Point Loss $\mathcal{L}_p$

Sec. 3 of the main paper describes the point-to-point loss $\mathcal{L}_p$ to penalize $Q_m \in \mathbf{Q}$ for being too far from $\mathbf{X}$. The point-to-point loss $\mathcal{L}_p$ is added to the loss function with the weight λ. Here, we show that our USIP is very robust to the value of λ. Specifically, the repeatability of our USIP keypoints remains almost the same over a wide range of values for λ. Keypoint repeatability is illustrated in Fig. 2 with various λ. Fig. 2 shows that the USIP keypoints are highly repeatable even when λ is small. This is probably because our design to limit the receptive field already guides the network to learn repeatable keypoints even without the point-to-point loss. On the other hand, the network fails to converge when λ is too large because the point-to-point loss dominates the training process. Nonetheless, training the network without the point-to-point loss does not ensure the keypoints to be close to the input point cloud. The top row of Fig. 1 shows keypoints from our USIP detector trained with $\lambda = 6$, *i.e.*, with point-to-point loss. They are close to the input point cloud. In comparison, the bottom row of Fig. 1 shows from our USIP detector trained without point-to-point loss, *i.e.*, $\lambda = 0$. These are less desirable keypoints that are farther from the input point cloud.

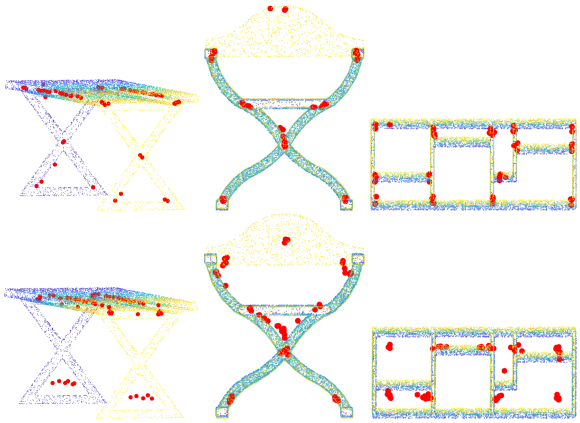


Figure 1. Visualization of USIP keypoints with different λ in Point-to-Point loss. First row $\lambda = 6$, second row $\lambda = 0$.

D. Our Descriptor a.k.a "Our Desc."

Fig 3 shows the network design of "Our Desc." inspired by 3DFeat-Net [6] as mentioned in Sec. 6.2 of the main paper. Given the output $(\mathbf{Q}, \mathbf{\Sigma})$ from FPN, a ball $\Omega_m(Q_m, r)$ of points from the point cloud $\mathbf{X}$ within a radius r is built around each $Q_m \in \mathbf{Q}$. A keypoint descriptor $f_m \in \mathbb{R}^L$ is extracted for each Ω_m. The descriptor can be trained with either weak [6] or strong supervision [7, 3]. We improve the keypoint descriptor training by utilizing the keypoint saliency uncertainty $\mathbf{\Sigma}$ in Sec. D.1, D.2, and E.

*Jiaxin Li now works at nuTonomy, an APTIV company.

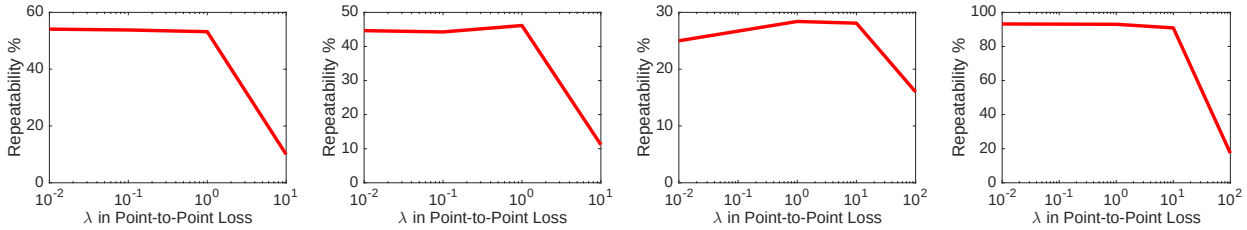


Figure 2. Relative repeatability with different weight λ for the Point-to-Point Loss $\mathcal{L}_p$. Number of keypoints is fixed to 128. Left to right: KITTI, Oxford, Redwood, ModelNet40.

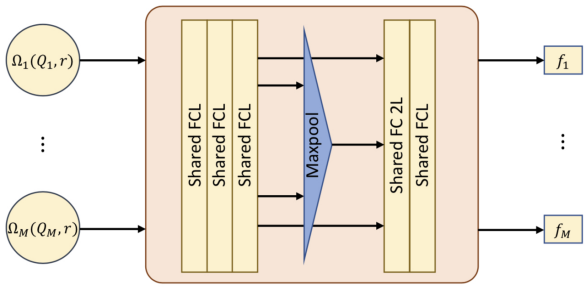


Figure 3. Network architecture of "Our Desc.".

D.1. Weak Supervision

Weak supervision of the descriptor is based on a triplet loss and the ground truth coarse registrations of the point clouds in the training dataset. Similar to [6], point clouds from the dataset are selected as the anchor samples during training. All overlapping pairs of point clouds to the anchor are defined as positive samples, while non-overlapping pairs of point clouds are defined as the negative samples. We denote the sets of keypoint descriptors extracted from the anchor, positive and negative samples as F_{anc}, F_{pos} and F_{neg}, respectively. We generate these training samples from the Oxford RobotCar and KITTI datasets. More formally, the triplet loss is given by:

$$\mathcal{L}_{\text{dc}}^{w} = \sum_{m=1}^{M} w_m \left\lfloor \min_{f_i \in F_{\text{pos}}} \|f_m - f_i\|_2 - \min_{f_j \in F_{\text{neg}}} \|f_m - f_j\|_2 + \gamma \right\rfloor_+, \tag{1}$$

where $f_m \in F_{\text{anc}}$ is a descriptor from the anchor sample. For each descriptor $f_m \in F_{\text{anc}}$, we minimize the Euclidean distance to its nearest neighbor $f_i \in F_{\text{pos}}$ and maximize the Euclidean distance to its nearest neighbor $f_j \in F_{\text{neg}}$. In addition, a normalized weight w_m is added to our triplet loss. w_m is derived from our USIP keypoint saliency uncertainty σ_m that indicates the reliability of Q_m and f_m. More specifically:

$$w_m = M \cdot \frac{\hat{w_m}}{\sum_{j=1}^{M} \hat{w_j}}, \qquad \hat{w_m} = \lfloor \xi - \sigma_m \rfloor_+, \tag{2}$$

where ξ is a threshold serves as the upper bound of σ_m.

D.2. Strong Supervision

We do strong supervision of the descriptor network on datasets with ground truth poses, *i.e.*, SceneNN [1] and "3D reconstruction dataset" [7]. The loss function for strong supervision defined on a pair of overlapping point clouds $\mathbf{X}$ and $\mathbf{X}'$ with ground truth poses $G \in \text{SE}(3)$ and $G' \in \text{SE}(3)$ is given by:

$$\mathcal{L}_{\text{dc}}^{s} = \sum_{m=1}^{M} w_m \left\lfloor \|f_m - f_i'\|_2 - \|f_m - f_j'\|_2 + \gamma \right\rfloor_+. \tag{3}$$

f_m and f_i' are keypoint descriptors from $\mathbf{X}$ and $\mathbf{X}'$, respectively. Additionally, f_i' is a descriptor with keypoint location Q_i' that is within a distance ρ from the keypoint location Q_m of the descriptor f_m, *i.e.*, $\|Q_m - GG'^{-1}Q_i'\|_2 < \rho$. To achieve hard negative mining, we randomly select 50% of f_j' from $\mathbf{X}'$ with the distance between the keypoint locations Q_j' and Q_m larger than ρ. The other 50% are chosen from keypoints with shortest but larger than ρ keypoint distances to Q_m.

E. More Point Cloud Registration Results

We follow the experimental setup and pipeline of 3DFeat-Net [6] to provide more evaluation results on point cloud registration. More specifically, we compare the performance of our USIP detector and "Our Desc." with other existing keypoint detector and descriptors. The evaluations are done on the Oxford RobotCar and KITTI datasets prepared by [6]. Refer to Sec. 6.2 of the main paper for the details of the registration steps. A fixed number of 256 keypoints is extracted from each point cloud. We extract the keypoints without Non-Maximum-Supression (NMS). Furthermore, keypoints with high saliency uncertainty, *i.e.*, large σ, are filtered out.

Datasets The Oxford RobotCar consists of 40 traversals on the same route over a year. 3D point clouds are built by accumulating the 2D scans from SICK LMS-151 LiDAR with the GPS/INS readings. We use 35 traversals, *i.e.* 21,875 point clouds for training. The remaining 5 traversals, *i.e.*, 828 point clouds and 3,426 overlapping pairs are

Method	Oxford					KITTI				
	RTE (m)	RRE (°)	Fail %	Inlier %	# Iter	RTE (m)	RRE (°)	Fail %	Inlier %	# Iter
ISS[8] + FPFH[4]	0.40 ± 0.29	1.60 ± 1.02	7.68	8.6	7171	0.33 ± 0.27	1.04 ± 0.77	39.00	8.8	8000
ISS[8] + SI[2]	0.42 ± 0.31	1.61 ± 1.12	12.55	4.7	9888	0.35 ± 0.31	1.11 ± 0.93	41.86	4.6	9401
ISS[8] + USC[5]	0.32 ± 0.27	1.22 ± 0.95	5.98	8.6	7084	0.27 ± 0.28	0.83 ± 0.76	18.62	7.7	8149
ISS[8] + CGF[3]	0.43 ± 0.32	1.62 ± 1.10	12.64	4.9	9628	0.23 ± 0.25	0.69 ± 0.60	8.90	8.4	7670
ISS[8] + 3DMatch[7]	0.49 ± 0.37	1.78 ± 1.21	30.94	5.4	9131	0.30 ± 0.28	0.80 ± 0.67	7.14	8.4	7165
3DFeat-Net[6]	0.30 ± 0.26	1.07 ± 0.85	1.90	13.7	2940	0.26 ± 0.26	0.56 ± 0.46	0.57	12.9	3768
USIP + Our Desc.	$\mathbf{0.28 \pm 0.26}$	$\mathbf{0.81 \pm 0.74}$	**0.93**	**28.1**	**523**	$\mathbf{0.21 \pm 0.24}$	$\mathbf{0.42 \pm 0.32}$	**0.24**	**28.0**	**600**

Table 1. Geometric registration performance on Oxford RobotCar and KITTI. The combination of our USIP keypoint detector and "Our Desc." outperforms existing methods in all criteria with around 2× inlier ratio.

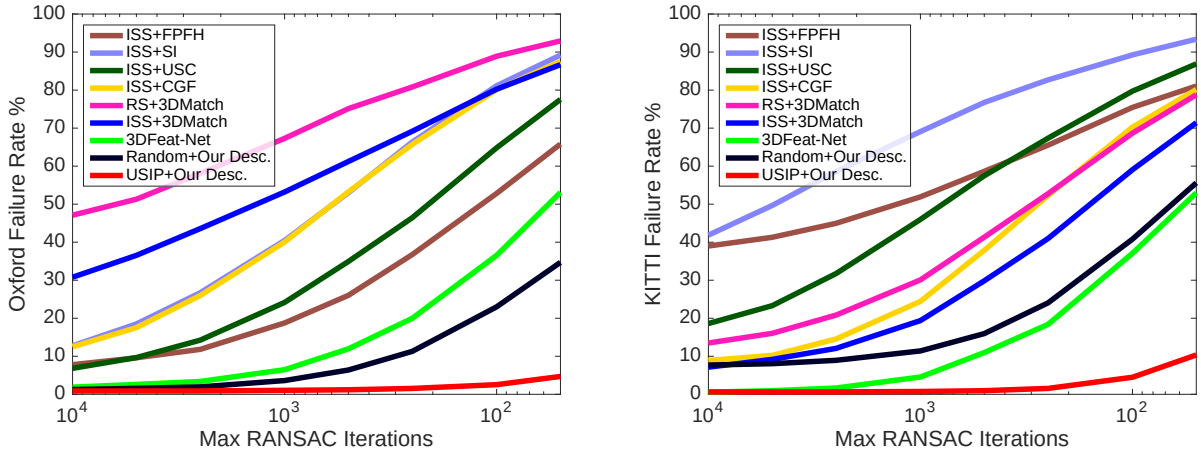
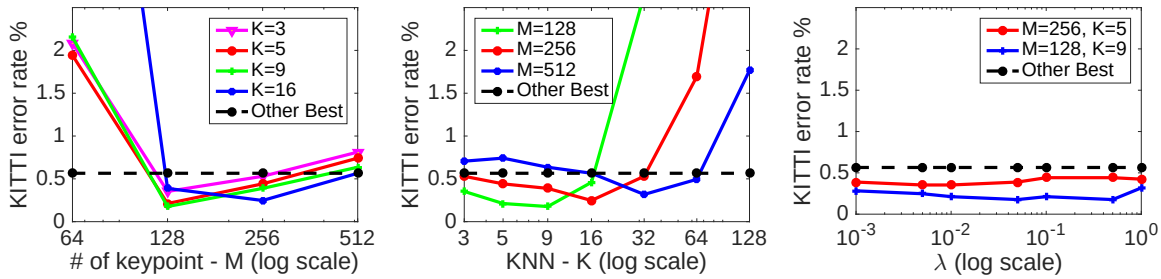


Figure 4. Registration failure rate versus maximum RANSAC iterations in Oxford RobotCar (left) and KITTI (right). Note that the x axis is in logarithmic scale. Our USIP detector + "Our Desc." (red line) shows very little drop in performance with decreasing number of RANSAC iterations.

used for evaluation. Random rotations around the up-axis are applied to each evaluation point cloud. In KITTI, 3D point clouds are directly provided by a Velodyne HDL-64E. We use the 2,831 overlapping pairs of point clouds prepared by [6] for registration evaluation.

Performance Tab. 1 shows the point cloud registration performances. Our USIP detector + "Our Desc." outperforms previous methods with the lowest registration failure rate (Fail %), Relative Translational Error (RTE), Relative Rotation Error (RRE), and highest inlier ratio (Inlier %). In particular, our registration failure rate and inlier ratio are respectively 50% and 2x of the second best keypoint detector + descriptor. We further analyze the performance over different number of RANSAC iterations. The registration failure rate versus the maximum number of RANSAC iterations is shown in Fig. 4. Due to high repeatability, our USIP detector (red line) shows very little drop in performance with decreasing number of RANSAC iterations, while all other algorithms show rapid drops in performances. Additionally, we replace our USIP detector + "Our Desc." with Random Sampling + "Our Desc." to demonstrate the effectiveness of our USIP detector. It can be seen from Fig. 4 that the performance of Random Sampling + "Our Desc." (black line) drops as quickly as other methods with decreasing number of RANSAC iterations.

Effect of USIP Keypoint Saliency Uncertainty Σ on Descriptor Training We show that the keypoint salicency uncertainty Σ from our USIP detector improves the performance of "Our Desc.". To this end, we compare the performances of "Our Desc." trained with USIP and randomly sampled keypoints, respectively. In particular, the weight w_m from Eq. 1 or Eq. 3 is set to 1 for the randomly sampled keypoints. We denote the descriptor trained with randomly sampled keypoints as "Desc. w. RS". Tab. 2 shows the registration failure rates of "Desc. w. USIP" and "Desc. w. RS". The results show that "Desc. w. USIP" performs better than "Desc. w. RS", which means that keypoints and saliency uncertainty Σ from our USIP detector improve descriptor training.



(a) (b) (c)

Figure 5. Point cloud registration error rate (%) on KITTI (trained on Oxford). Dash line is the best performance of existing methods. $\lambda = 0.5$ in (a) (b).

Failure %	Oxford		KITTI	
	Desc w. USIP	Desc w. RS	Desc w. USIP	Desc w. RS
USIP	**0.93**	1.20	**0.24**	1.02

Table 2. Registration failure rate for "Our Desc." trained keypoints from our USIP detector and randomly sampled keypoints.

Effect of Parameters M, K, λ We demonstrate the point cloud registration failure rate (%) in Fig. 5, when various USIP detector parameters, M, K, λ, are selected. In Fig. 5 we use the same descriptor mentioned in Sec. D. As shown in Fig. 5, our method outperforms existing methods over a wide range of M, K, λ. We notice our network performance decreases significantly when M is too small or K is too large, *i.e.*, the receptive is too large. This further verifies our design of limiting the receptive field. In addition, Fig. 5 shows that the registration failure rate remains satisfying when λ is small. This is consistent with Fig. 2 that our USIP is able to detect repeatable keypoints even without the point-to-point loss. Nonetheless, it is still important to include the point-to-point loss to ensure that the keypoints are close to the input point cloud.

F. Qualitative Visualization of USIP Keypoints

We show more visualizations of the keypoints detected from our USIP detector on ModelNet40, KITTI, Oxford RobotCar, Redwood in Fig. 8, 9, 10, 11, respectively. NMS and Σ thresholding are applied here. A limitation of our USIP detector is shown in Fig. 8, where there are no or very few keypoints on objects that are highly symmetrical or with smooth surfaces. The saliency uncertainties Σ of the keypoints detected on these objects are large, thus discarded by the Σ thresholding.

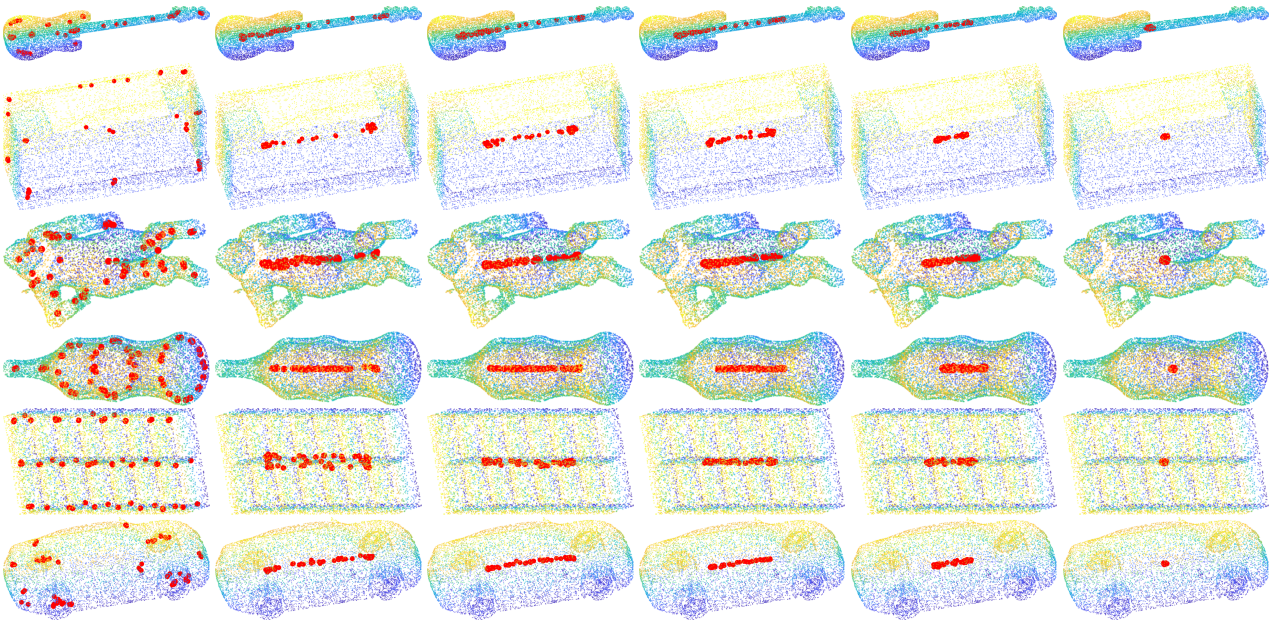


Figure 6. Visualization of FPN degeneracy. $M = 64$ and from left to right: $K = 9, 24, 32, 40, 48, 64$, *i.e.*, receptive field of FPN increases from left to right.

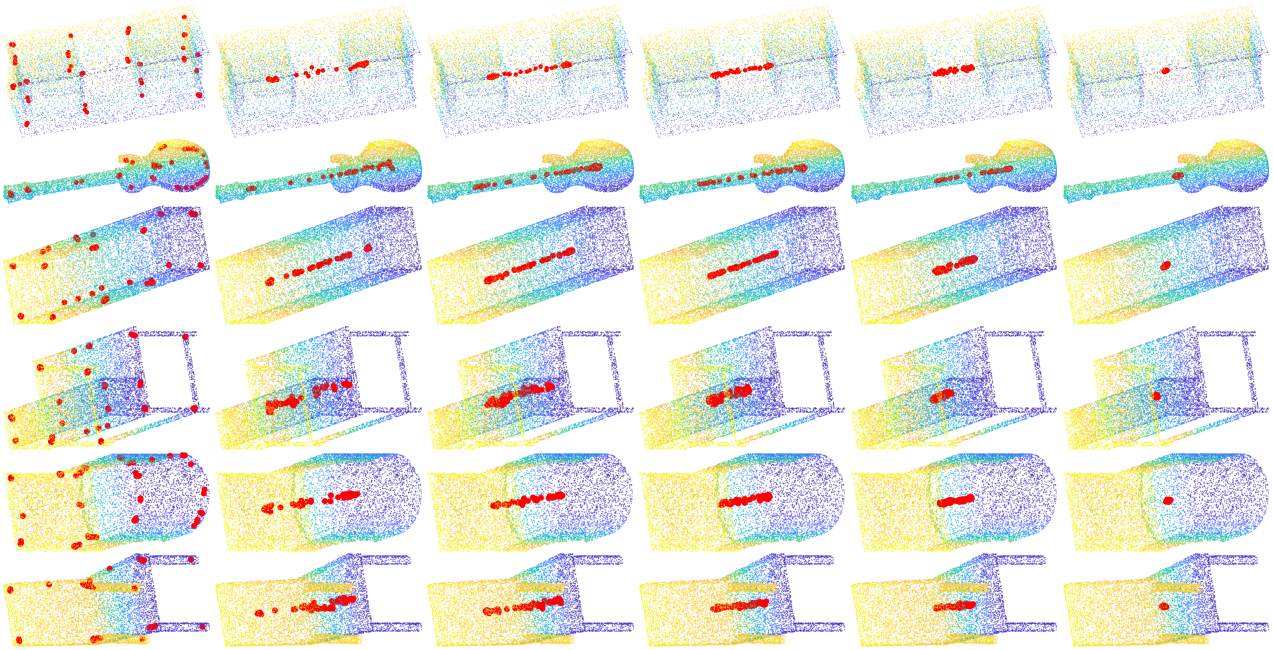


Figure 7. Visualization of FPN degeneracy. $K = 9$ and from left to right: $M = 64, 24, 20, 16, 12, 9$, *i.e.*, receptive field of FPN increases from left to right.

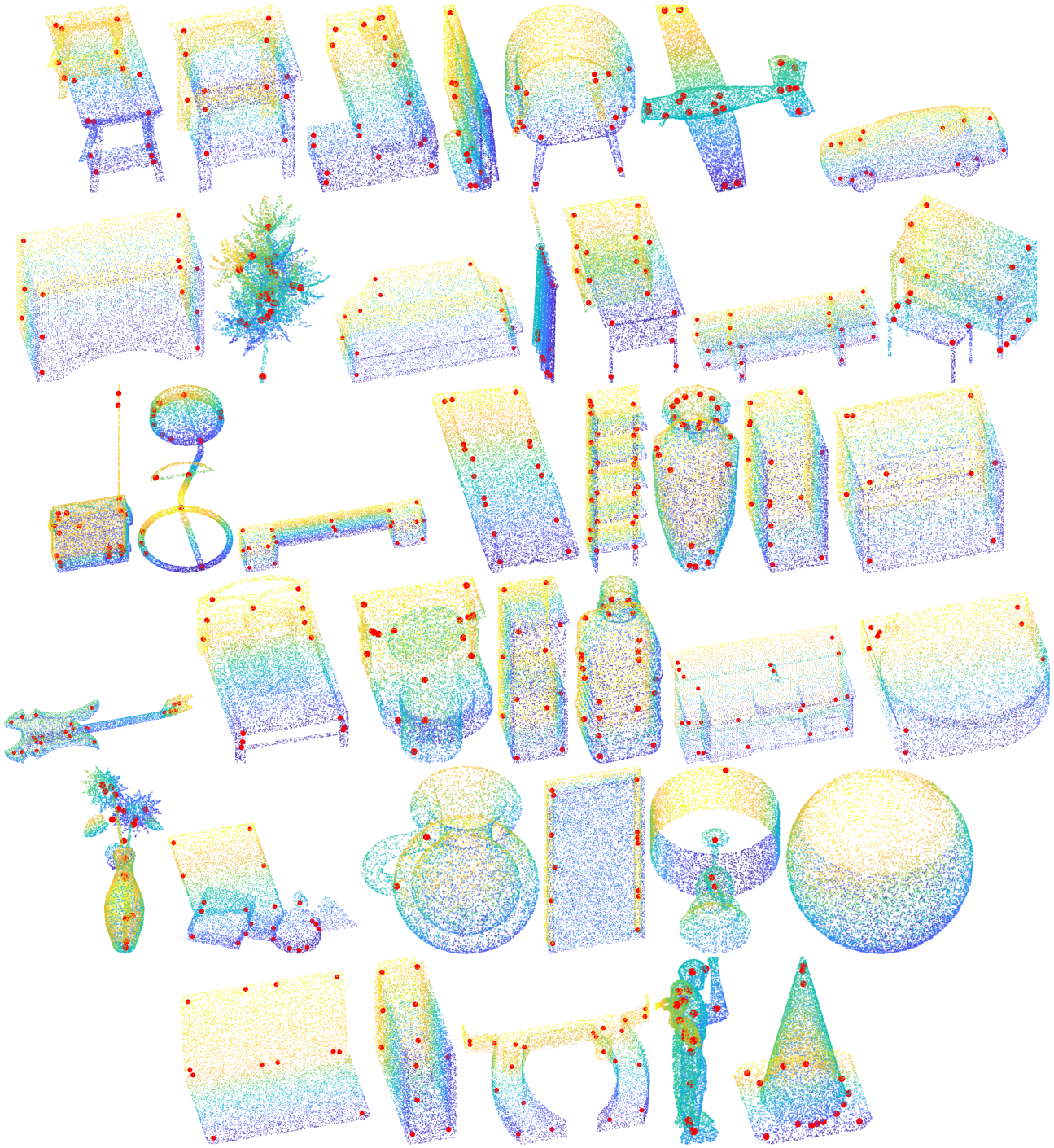


Figure 8. Visualization of USIP keypoints on ModelNet40. Best view with color and zoom-in.

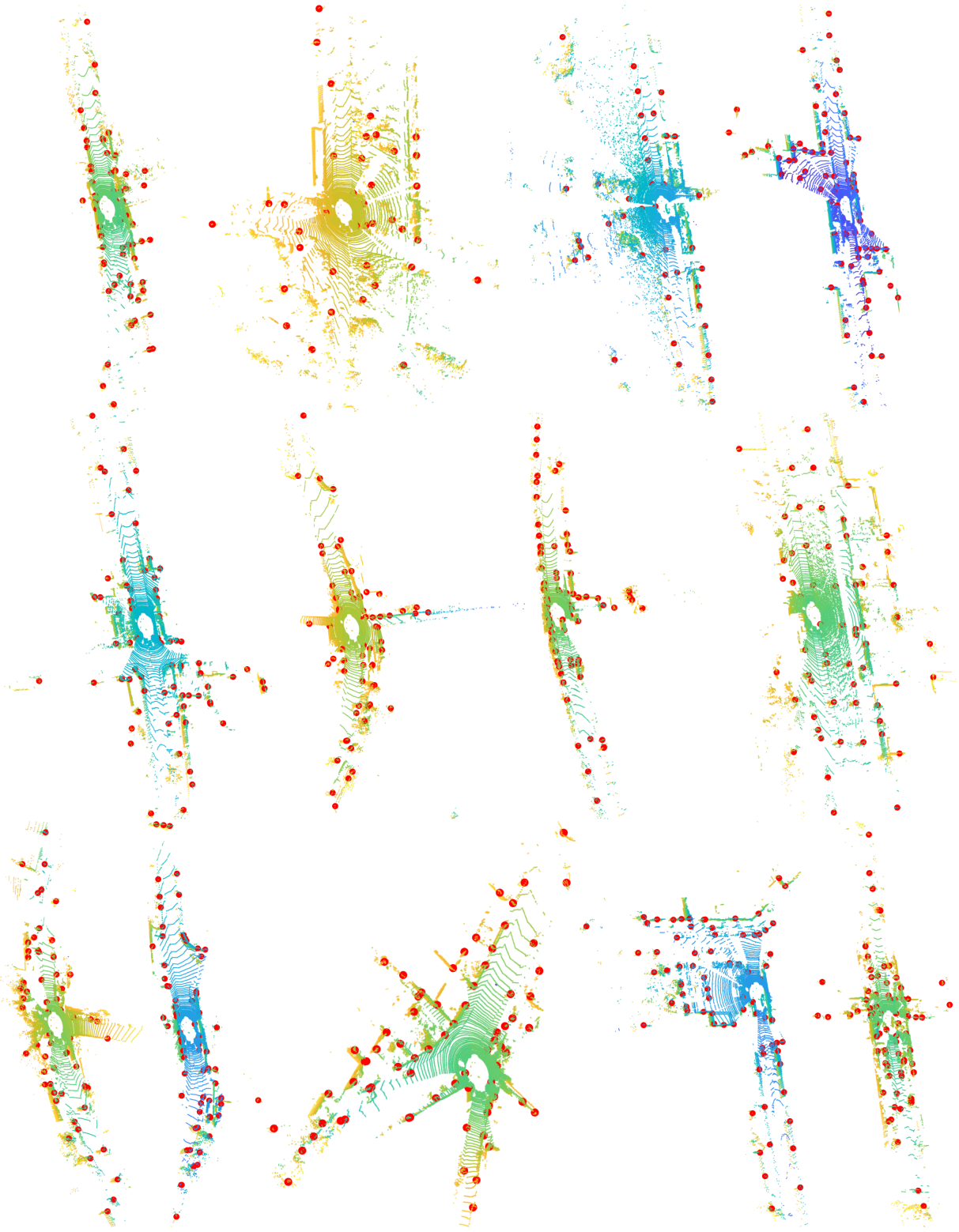


Figure 9. Visualization of USIP keypoints on KITTI with our USIP detector trained on Oxford RobotCar dataset. Best view with color and zoom-in.

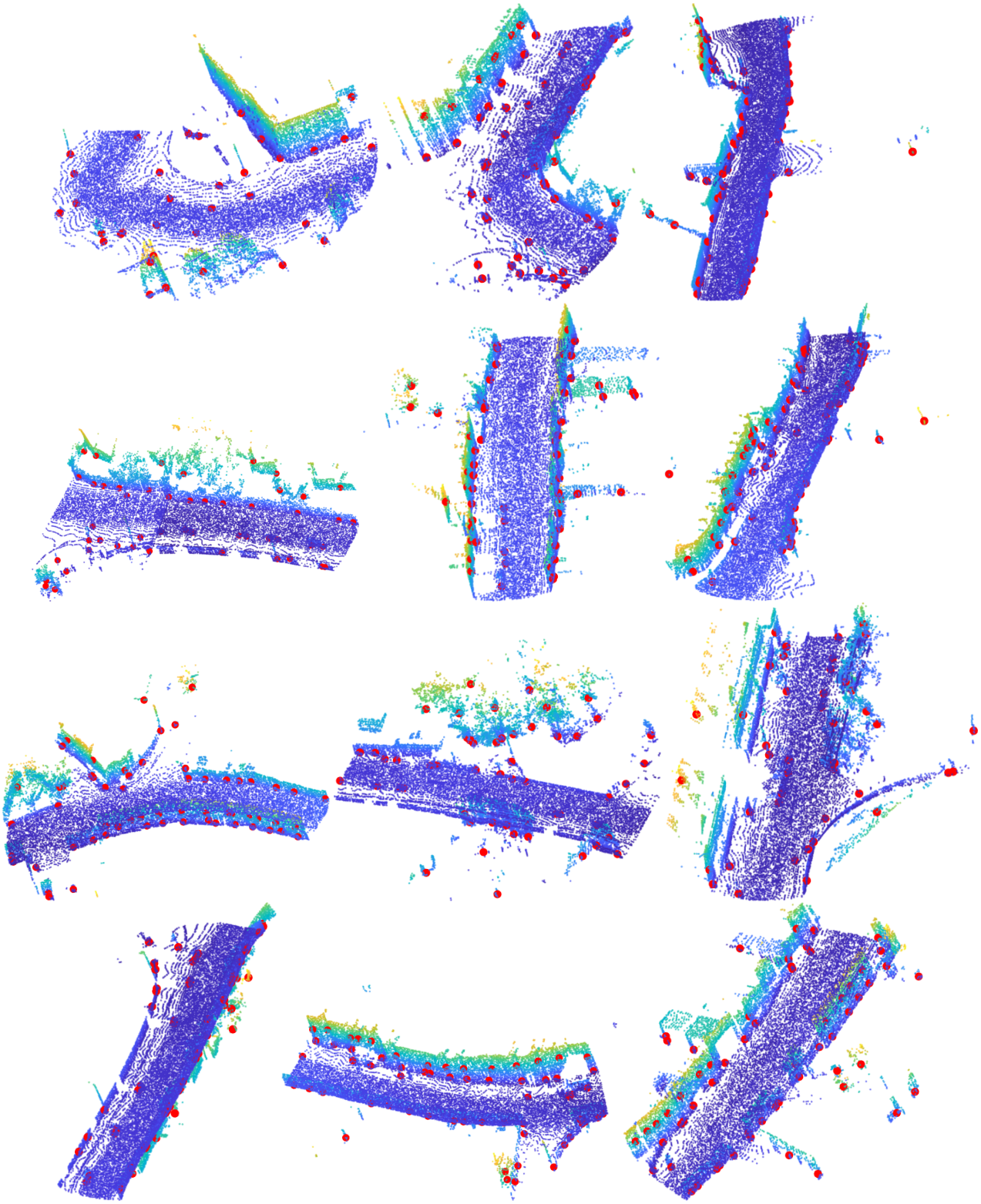


Figure 10. Visualization of USIP keypoints on Oxford RobotCar. Best view with color and zoom-in.

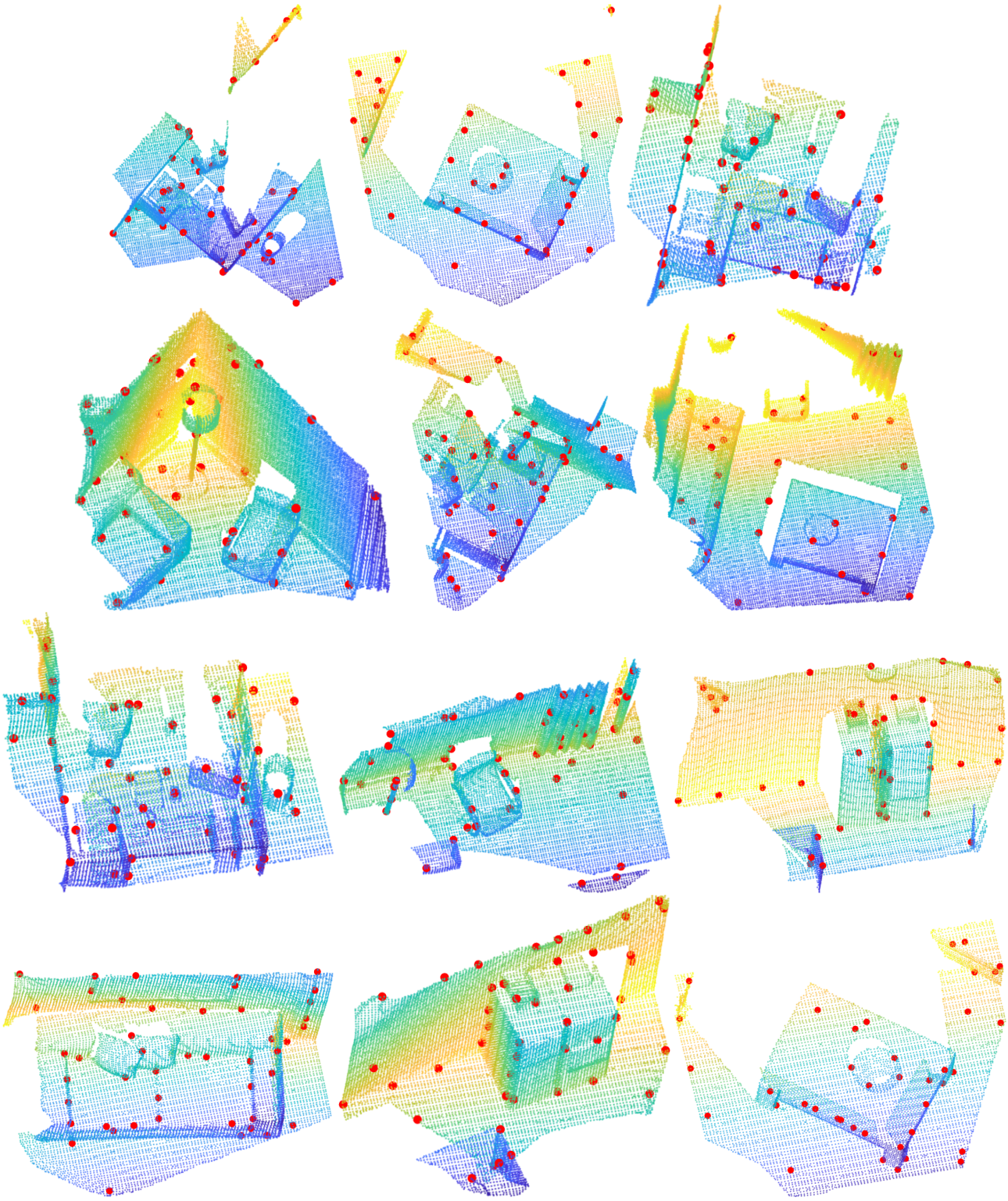


Figure 11. Visualization of USIP keypoints on Redwood with our USIP detector trained on "3D Reconstruction Dataset" [7]. Best view with color and zoom-in.

References

[1] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 92–101. IEEE, 2016. 2

[2] Andrew E Johnson and Martial Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):433–449, 1999. 3

[3] Marc Khoury, Qian-Yi Zhou, and Vladlen Koltun. Learning compact geometric features. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 153–61, 2017. 1, 3

[4] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 3212–3217. Citeseer, 2009. 3

[5] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique shape context for 3d data description. In *Proceedings of the ACM workshop on 3D object retrieval*, pages 57–62. ACM, 2010. 3

[6] Zi Jian Yew and Gim Hee Lee. 3dfeat-net: Weakly supervised local 3d features for point cloud registration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 607–623, 2018. 1, 2, 3

[7] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 199–208. IEEE, 2017. 1, 2, 3, 9

[8] Yu Zhong. Intrinsic shape signatures: A shape descriptor for 3d object recognition. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 689–696. IEEE, 2009. 3