Supplementary Material *for* SC3K: Self-supervised and Coherent 3D Keypoints Estimation from Rotated, Noisy, and Decimated Point Cloud Data

Mohammad Zohaib, Alessio Del Bue
Pattern Analysis & Computer Vision (PAVIS)
Italian Institute of Technology (IIT), Genoa, Italy
{mohammad.zohaib, alessio.delbue}@iit.it

1. Introduction

This file contains supplementary material for *SC3K: Self-supervised and Coherent 3D Keypoints Estimation from Rotated, Noisy, and Decimated Point Cloud Data*. Due to the limited space in the main paper, here we present additional ablations related to the experimental section of the paper, provide a complete table (as given in [4]) comparing the DAS of our approach with other methods, and show the qualitative results of our experiments. Moreover, we also share a video representing 3D visualizations of the results reported in the main paper (i.e., the keypoints estimated by SC3K and SOTA approaches). The video (rotating keypoints) makes it easy to understand the robustness of the SC3K.

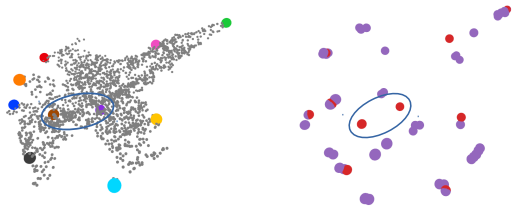
2. Additional Ablations

This section presents seven additional ablations: 1) comparison of the estimated keypoints with the vertices of a convex hull (CH) of the object, 2) evaluation of SC3K on partial PCDs, 3) performance of SC3K for different number of keypoints, 4) sensitivity of the inclusivity metric with respect to (w.r.t.) the threshold τ_2, 5) impact of the separation and shape loss on the performance of SC3K, 6) impact of the residual block on SC3K, and 7) evaluation of SC3K for different augmentations.

2.1. Keypoints vs. vertices of a Convex Hull (CH)

Since our approach estimates keypoints on the object's surface, one can relate them with the vertices of the CH of the object. Therefore in this section, we highlight the significance of our keypoints over the vertices of the CH. We observed that the vertices of the CH are not as expressive as our semantic keypoints due to the following reasons:

- CH vertices do not maintain semantic information (ordering) that represents correspondences between two or more views. To validate this, we implemented a new baseline where we computed vertices of the CH of the test samples, selected 10 meaningful vertices (since 10 keypoints are used for comparison) by using K-means clustering, and computed the DAS metric for them. It is found that the DAS of SC3K (82.86) is **+68.87** higher than the one obtained from the CH vertices (13.99).
- CH vertices will not include points that lie within the convex hull (see Fig. 1), which might be important for some downstream tasks, e.g. characterising the shape of objects.



(a) Keypoints of SC3K (b) Vertices of the CH

Figure 1: Comparison: (a) keypoints estimated by SC3K w.r.t. the object, (b) vertices of the CH w.r.t. the keypoints of the SC3K. Some of the keypoints are not included in the vertices of the CH.

2.2. Evaluation od SC3K on partial PCDs

SC3K is flexible in the backbone used, so it is possible to insert a Point Completion Network (PCN) [5] that accepts partial PCDs. The updated network first estimates the missing parts of the object and then predicts the keypoints. We use the pretrianed weights of the PCN. We consider common categories of the PCN

Data type	Inclusivity	Coverage	DAS	Average
Seen categories				
Partial	46.65	89.12	54.46	63.41
Dense	77.61	94.66	79.45	83.90
Complete	**90.69**	**95.09**	**80.27**	**88.68**
Unseen categories				
Partial	62.78	91.40	51.92	68.70
Dense	48.27	94.11	**72.73**	71.70
Complete	**87.11**	**96.76**	64.27	**82.71**

Table 1: Results for the PCN dataset. We test SC3K separately on PCN seen and unseen categories. On average, the results are improved when the input shapes are completed (dense) using the PCN network.

and keypointNet dataset and select five seen (airplane, car, chair, table, vessel) and three unseen (bed, guitar, motorbike) categories. The results for the selected categories are given in Tab. 1. As expected, the performance is higher for estimated complete PCDs (called as dense) and lower for partial PCDs. Consider that the updated approach depends on the shape completion module. Thus results are not very impressive where the shapes are not properly estimated. Fig. 2. shows keypoints predicted on partial and estimated complete PCDs.

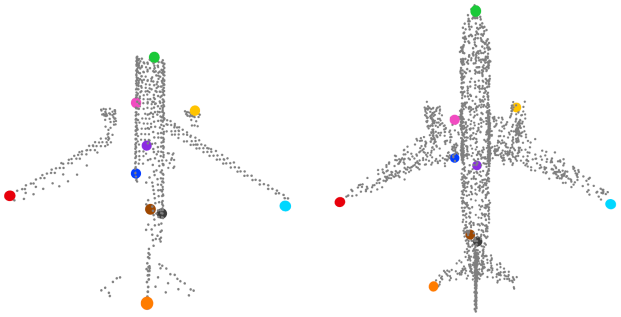


Figure 2: Visualizations of the keypoints estimated on partial and completed PCD.

2.3. Performance of the proposed approach for different numbers of keypoints

We evaluate our approach by varying the number of computed keypoints from the PCD. We found that for most of the shapes (e.g., bottle, guitar), our approach estimates keypoints over the surface of the object. However, for the detailed objects with gaps between the parts (i.e., airplanes have relevant empty spaces between a wing and the tail), some of the keypoints are estimated outside the object (in the gaps). This effect appears only when a high number of keypoints are considered (higher than 35). As an example, different numbers of keypoints estimated for the cup and airplane category are shown in Fig. 3.

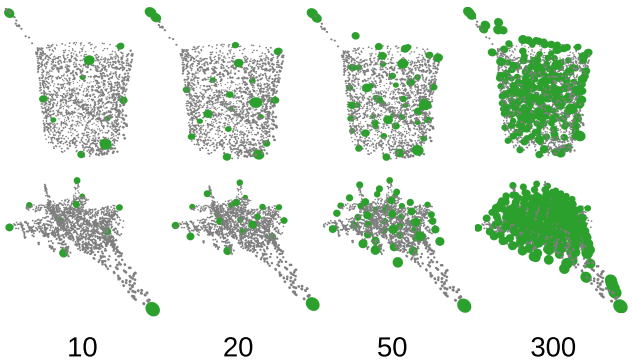


Figure 3: Estimation of different numbers of keypoints for the same object. The keypoints are estimated on the object's surface if they are less than or equal to 35 in number. They are predicted outside the object (in case of more than 35 keypoints), especially for a detailed object having empty spaces among its parts.

2.4. Inclusivity metric and its sensitivity w.r.t. to the parameter τ_2

The inclusivity metric (defined by [1]) depends on the total number of keypoints and the tolerance threshold τ_2. To validate this, we train our network separately for different numbers of keypoints, and calculate the inclusivity for different τ_2. It is found that the inclusivity is higher for fewer keypoints, and it increases with the increase in the τ_2. Fig. 4 shows the average inclusivity (of the test set) for different values of τ_2.

2.5. Impact of the separation and shape loss on perturbation

We observed that the separation and shape loss contribute to robustness, especially in the case of perturbation. To validate this, we have trained SC3K (for airplane category) without separation and shape loss and tested it for noisy and down-sampled PCDs. The results are illustrated in Tab. 2. The performance drops significantly with an increase in the noise ratio or down-sampling scale.

2.6. Impact of residual blocks

We evaluated SC3K by replacing the residual blocks with Conv1D layers obtaining a performance decrease of -3.54% -8.21% -6.70% on inclusivity, convergence and DAS on keypointNet dataset. Such a remarkable drop supports our design choice.

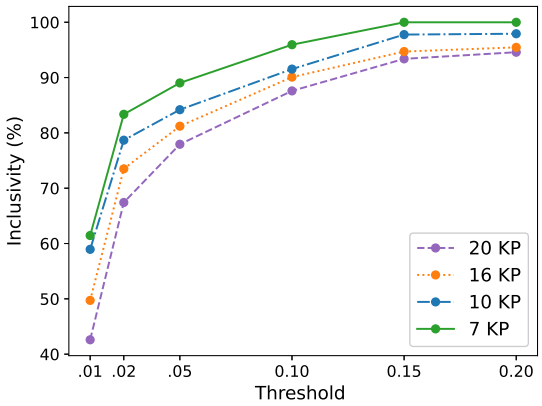


Figure 4: Average inclusivity of the proposed approach for different keypoints and thresholds (τ_2). The inclusivity increases with an increase in the τ_2, and it is higher for fewer keypoints.

Noise/down-sample scale		0/0x	0.02/4x	0.06/16x
All loss	Inc.	87.2/87.2	83.1/84.9	79.6/82.6
	Cov.	96.3/96.3	95.2/96.1	92.6/95.6
Without two losses	Inc.	80.2/80.2	74.9/60.1	59.2/50.8
	Cov.	90.9/90.9	87.8/86.0	84.5/82.2

Table 2: Results for PCD perturbations without using separation and shape loss

2.7. Evaluation for different augmentation strategies

As reported in Tab. 3, we evaluate the effect of different augmentations during training, such as downsample (DS), Noise (N) and Rotation (R). We used the same test set (canonical, without noise and down-sampling) for evaluation. Tab. 3 shows that the coverage and inclusivity decrease due to augmentations, whereas DAS increases when we use noisy and downsampled train samples. We observed that for noisy downsampled samples, although the keypoints are semantically consistent, they are not estimated on the object's surface. Considering the average value, we suggest that only rotations should be used as augmentation for training.

Augmentation	Inclusivity	Coverage	DAS	Average
R	**75.89**	**95.63**	69.71	**80.41**
R + DS	73.61	89.56	71.72	78.30
R + N	70.78	91.11	65.98	75.95
R + DS + N	70.47	89.68	**77.68**	79.27

Table 3: Evaluation of SC3K for different augmentations.

3. Quantitative results

This section presents some additional quantitative comparisons of SC3K with SOTA approaches.

3.1. Comparison with USEEK [3]

We compare our keypoints for random-oriented PCDs with those of USEEK. We select the four categories (airplane, chair, guitar and knife) for which the USEEK's pretrained weights are available. The results are reported in Tab. 4. We found that, on average, SC3K's inclusivity and coverage are +16.71% and +23.82% higher than those of USEEK. However, USEEK estimates only 3 and 4 keypoints for knife and guitar, respectively, so those are well separated and semantically ordered. Moreover, considering pose estimation as a downstream task, we use the keypoints estimated by both methods to compute a relative pose between two randomly oriented PCDs. The mean/median of the pose error of SC3K is +19.75°/+5.69° better than that of USEEK. Thanks to the mutual loss components for enabling SC3K to estimate aligned keypoints irrespective of orientation.

Approach	Inclusivity ↑	Coverage ↑	DAS ↑	Pose error ↓ Mean	Pose error ↓ Median
USEEK	80.49	73.42	77.78	30.12	6.45
SC3K	97.20	97.24	72.70	10.72	0.76
Difference	+16.71	+23.82	-5.08	+19.40	+5.69

Table 4: Comparison of SC3K with USEEK [3]. The keypoints estimated by SC3K are comparatively more useful for computing relative pose between two transformed versions of the same object.

3.2. Comparison with ISS [6] and MR [4]

In this section, we compare DAS of SC3K with the other baseline approaches, i.e., ISS [6] and MR [4]. We consider their DAS exactly the same as reported in [4]. Our results are the same as we have reported in the main paper. However, we only show the DAS of the same categories as given in [4]. It can be observed that, on average, SC3K outperforms the other approaches.

4. Qualitative results

This section presents a qualitative comparison of the SC3K with the SOTA approaches. Moreover, it also shows the keypoints estimated by SC3K for intra-class, noisy and down-sampled objects.

4.1. Comparison with ULCS [1] and SM [2]

We show in this section the keypoints estimated by SC3K and the SOTA approaches in Fig. 5. For better

	ULCS [1]	SM [2]	ISS [6]	MR [4]	SC3K
Airplane	61.40	77.70	13.10	81.00	**82.86**
Chair	64.30	76.80	10.70	83.10	**87.04**
Car	–	**79.40**	8.00	74.00	75.19
Table	–	70.00	16.20	**78.50**	76.03
Guitar	–	63.10	8.70	61.30	**65.67**
Mug	–	67.20	11.20	68.20	**79.25**
Cap	–	53.00	13.10	57.10	**59.72**
Mean	62.85	69.60	11.57	71.89	**74.54**

Table 5: Comparison based on the semantic consistency between the keypoints estimated for different objects of the same category. The baseline results (DAS) are the same as reported in [4]. The higher value is best.

understanding, the estimated keypoints (in different colours) are shown on top of the original PCDs (in Gray). The colour of the keypoints represents their semantic ID information, i.e. a point with the same colour should stay in the same area despite perturbations. Columns 1 and 2 illustrate the keypoints estimated by ULCS [1] and SM [2], respectively. In contrast, two views of the keypoints estimated by the proposed approach are depicted in columns 3 and 4. The comparison validates that our keypoints are estimated close to the surface, highlighting the corners, thus best characterizing the object's shape.

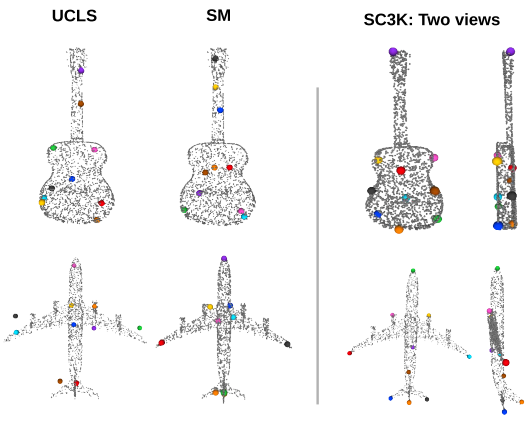


Figure 5: Qualitative comparison. Columns 1 and 2 present keypoints estimated by ULCS and SM, respectively. Columns 3 and 4 show the keypoints estimated by SC3K. It can be observed that some of keypoints of the ULCS are estimated outside the object (airplane). The keypoints estimated by SC3K best characterize the object's shape, as they are estimated on the surface and cover the complete object.

4.2. Qualitative comparison with Intra-class objects

This section shows the qualitative results of the SC3K for intra-class objects. Four objects (in random poses) of the different categories are shown in Fig. 6. It can be observed that the keypoints are proximal to the original PCDs, semantically in order (coherent) and pointing to the sharp edges of the objects.

4.3. Visualisation of the noisy PCDs

This section shows the qualitative results (extension of the Fig. 5a in the main paper) of SC3K for different noisy PCDs. We add the Gaussian noise of different scales to the original PCDs of different categories. The noise scale is written in the beginning of every row where "0.00" mean original PCD without noise. The estimated keypoints are shown in Fig. 7. It can be observed that the proposed SC3K remains successful in estimating the 3D keypoints from the noisy PCDs. Moreover, the keypoints are always estimated close to the outermost points in the PCDs (i.e. close to the noisy surface). However, the accuracy decreases with the increase in the noise scale.

4.4. Visualisation of the Down-sampled PCDs

This section presents the performance of SC3K for down-sampled PCDs as an extension of the results shown in the Fig. 5b of the main paper. For decimating the PCD, we use the Farthest Point Sampling (FPS) as used in [5] to sample points from original PCDs for different sampling ratios. We test our pre-trained network to estimate the 3D keypoints from the down-sampled PCDs. The results are shown in Fig. 8. The figure is horizontally divided to fix all the objects on one page. Each column presents the results of a different object. The sampling ratio is shown at the beginning of every row. The "0×" shows the original PCD without sampling (zero times sampling). It can be observed that the SC3K has estimated approximately accurate keypoints for the down-sampled PCDs. However, the keypoints are not estimated at the same positions as the positions of the corresponding keypoints of the original PCDs (without sampling) when the PCDs are scaled 32 times (32×). The 32× sampling means a PCD containing only 64 points, considering that the original PCD contains 2048 points.

References

[1] Clara Fernandez-Labrador, Ajad Chhatkuli, Danda Pani Paudel, Jose J Guerrero, Cédric Demonceaux, and Luc Van Gool. Unsupervised learning of category-specific symmetric 3d keypoints from point sets. In *European Conference on Computer Vision*, pages 546–563. Springer, 2020. 2, 3, 4

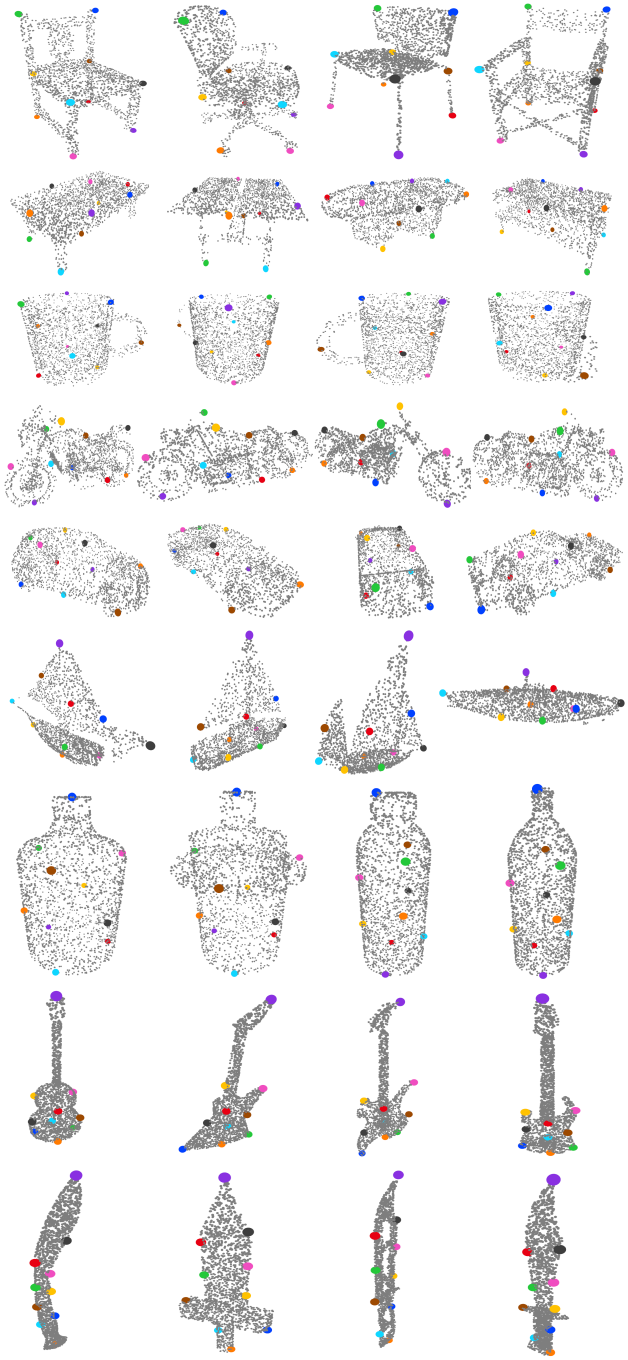


Figure 6: Qualitative results of the proposed SE3K for different categories. Every row shows four objects (in different poses) of the same category. The keypoints (coloured points) are estimated on the surface and in the same pose as the pose of the original PCDs (small gray points). Moreover, they are semantically consistent for all the intra-class objects.

[2] Ruoxi Shi, Zhengrong Xue, Yang You, and Cewu Lu. Skeleton merger: an unsupervised aligned keypoint detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 43–52, 2021. 3, 4

[3] Zhengrong Xue, Zhecheng Yuan, Jiashun Wang, Xueqian Wang, Yang Gao, and Huazhe Xu. Useek: Unsupervised se(3)-equivariant 3d keypoints for generalizable manipulation. 2023. 3

[4] Haocheng Yuan, Chen Zhao, Shichao Fan, Jiaxi Jiang, and Jiaqi Yang. Unsupervised learning of 3d semantic keypoints with mutual reconstruction. *arXiv preprint arXiv:2203.10212*, 2022. 1, 3, 4

[5] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *International Conference on 3D Vision*, pages 728–737. IEEE, 2018. 1, 4

[6] Yu Zhong. Intrinsic shape signatures: A shape descriptor for 3d object recognition. In *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*, pages 689–696. IEEE, 2009. 3, 4

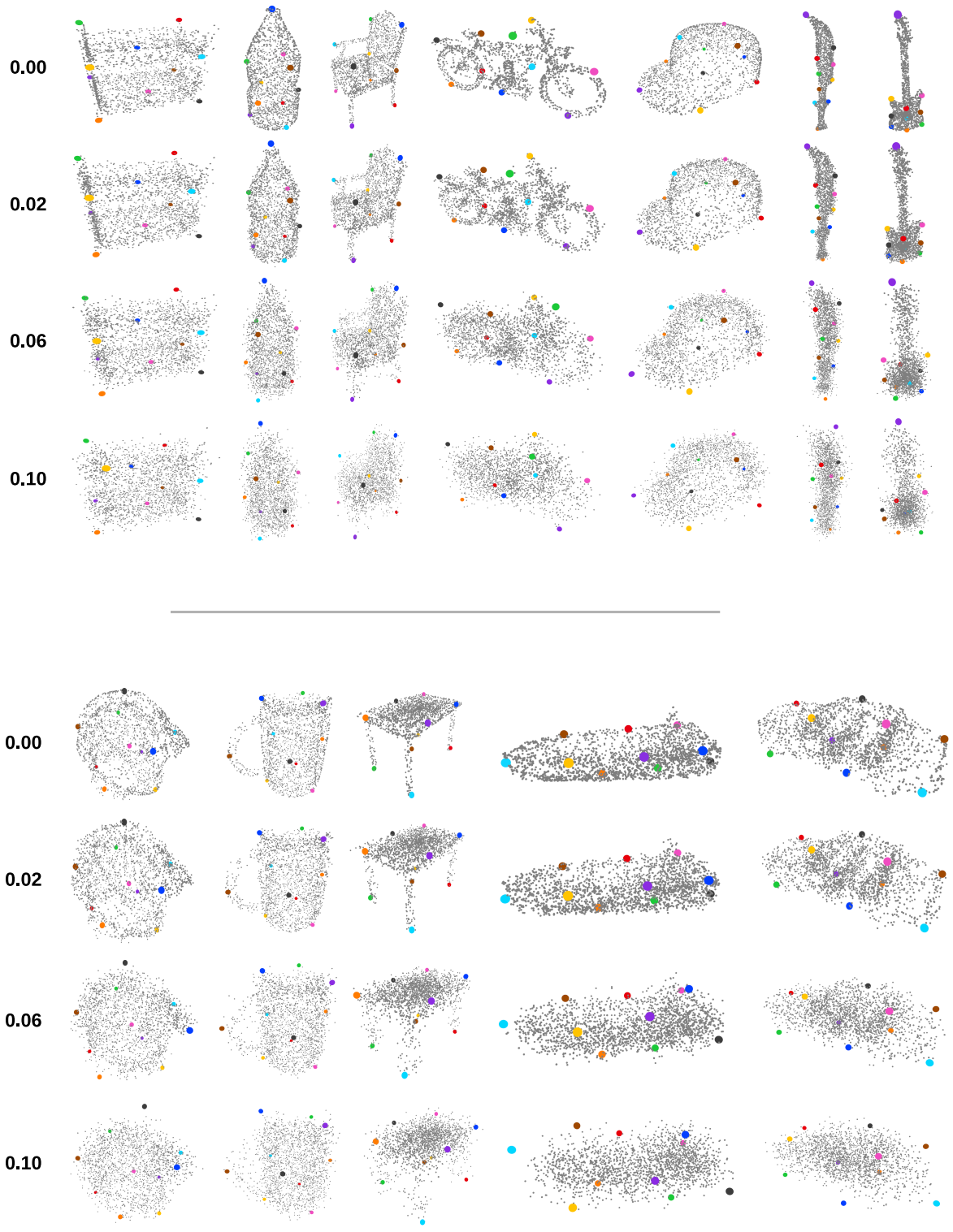


Figure 7: Performance of the proposed approach for the noisy PCDs. Gaussian noise of different scales (as mentioned at the beginning of every row) is added to the input PCDs. "0.00" represents the original PCD (without noise). The SC3K remains successful in estimating the semantically consistent keypoints for noisy PCDs. However, the accuracy has decreased with an increase in the noise scale.

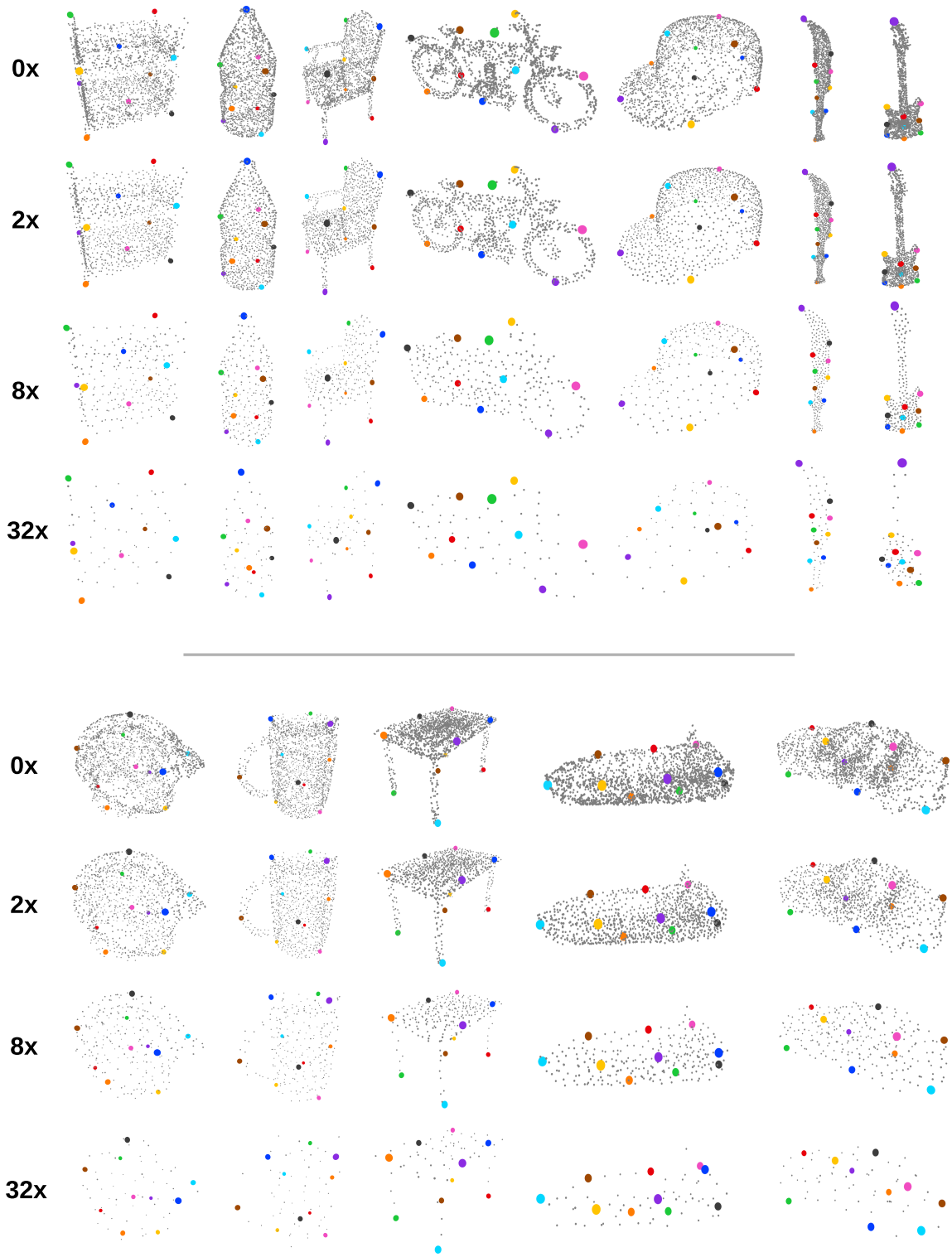


Figure 8: Performance of our method for down-sampled PCDs. The input PCDs are down-sampled for different scales, as mentioned at the beginning of every row. The "0×" shows the original PCDs. The proposed SC3K remains successful in estimating the approximately accurate 3D positions of the keypoints.