SGPA: Structure-Guided Prior Adaptation for Category-Level 6D Object Pose Estimation (Supplementary Material)

Kai Chen and Qi Dou
Department of Computer Science and Engineering, The Chinese University of Hong Kong
{kaichen, qdou}@cse.cuhk.edu.hk

Overview

In this supplementary material, we provide additional contents of SGPA network that are not included in the main paper due to the space limit:

- Section A gives a more detailed per-category pose accuracy comparison between SGPA and SPD on the REAL275 dataset.
- Section B provides more qualitative comparisons with SPD.
- Section C shows more visualization results of the learned attention map on the REAL25 dataset.
- Section D shows the key-points extracted by our structure regularized low-rank transformer network. Section E deploys our prior adaptation module on SPD. The comparative results further demonstrate the effectiveness of our proposed prior adaptation network.

A. Per-Category Evaluation

Figure 1 and Figure2 show detailed comparisons between our SGPA and SPD [1] on the REAL275 dataset. SGPA significantly outperforms SPD in terms of rotation, especially for object categories that may contain a large intra-class variasion in shape and size, such as *camera* and *mug*. For the camera category, its lens could be very different from one instance to another in length. For the mug category, ite overall size and the shape of its handle could have obvious variations. On these categories, our SGPA achieves a much more higher accuracy than SPD, which demonstrates the superiority of our method.

B. Qualitative Results on REAL275

Figure 3 gives more qualitative comparisons between SGPA and SPD on the REAL275 dataset [2]. We visualize the pose estimation result by displaying the predicted object axes on the RGB image. Ideally, the center and the orientation of the axes should be well aligned with the object (as shown in Figure 3(a)). In different real environments, our proposed SGPA consistently outperforms SPD, which can be best viewed and demonstrated by comparing the orientation of object axes predicted by different methods.

Table 1. **General effectiveness of proposed prior adaptation (Section E).** Evaluation of the proposed prior adaptation method when integrated with SPD.

Method	CAMERA25					
	$3D_{50}$	$3D_{75}$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$
SPD*[1]	93.0	85.5	58.1	62.9	75.9	83.8
+ Adaptation	**93.2**	**87.5**	**64.9**	**69.5**	**78.8**	**85.7**
	+0.2	+2.0	+6.8	+6.6	+3.9	+1.9
	REAL275					
	$3D_{50}$	$3D_{75}$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$
SPD*[1]	80.0	56.7	20.0	22.3	45.3	57.9
+ Adaptation	**80.8**	**58.8**	**26.5**	**30.0**	**53.1**	**63.4**
	+0.3	+2.1	+6.5	+7.7	+7.8	+5.5

C. Visualization of the Attention Map

Figure 4 presents more visualization results of the attention map learned by our SGPA network. The learned attention map indicates the relationship (structure similarity) between the prior point cloud and the object point cloud. For a query position on the prior point cloud, we visualize its associated relationships with the object point cloud by projecting the attention values onto the corresponding RGB image. For each attention map, the warmer of the color, the larger of the attention value. Red indicates a stronger attention. For each query position of the prior point cloud, our SGPA tends to first focus on the corresponding part of the instance (e.g., see the point on the handle of the mug and its top-8 attention map), and then spread to the whole object region to learn a global relationship. These results demonstrate that SGPA learns meaningful structure similarity between prior and target object for effective prior adaptation.

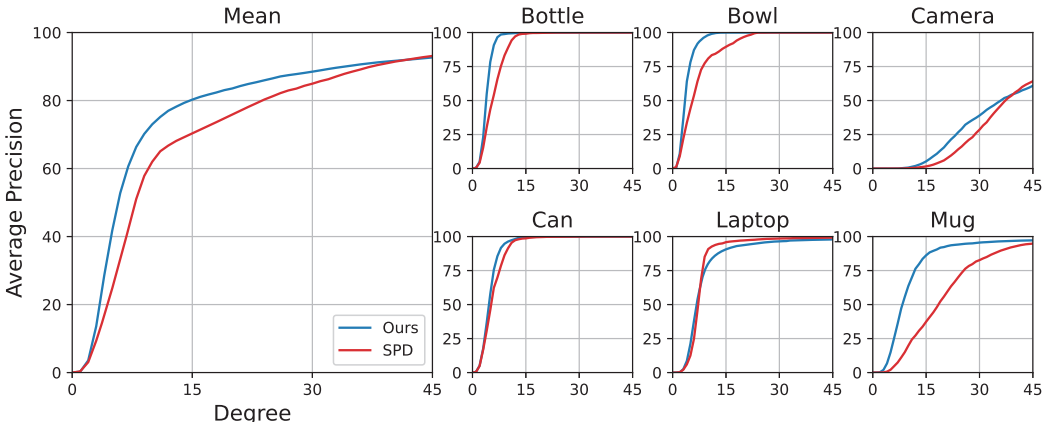


Figure 1. **Rotation precision evaluation (Section A).** We give a detailed comparison of our SGPA and SPD [1] on the REAL275 dataset in terms of rotation accuracy. The average precision with respect to different angle thresholds are computed and compared.

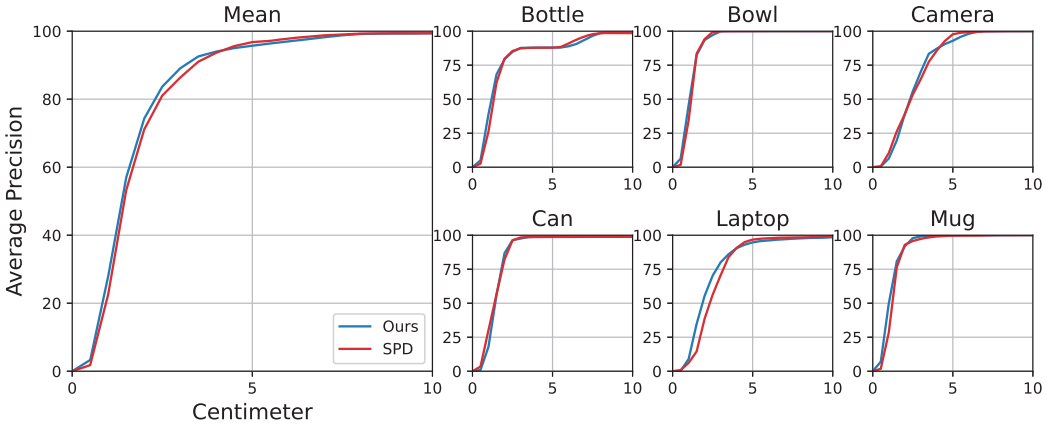


Figure 2. **Translation precision evaluation (Section A).** We give a detailed comparison of our SGPA and SPD [1] on the REAL275 dataset in terms of translation accuracy. The average precision with respect to different translation thresholds are computed and compared.

D. Visualization of Object Key-Points

In Section 3.3 of the main paper, we resort to an auxiliary network to extract n key-points from the target object point cloud, and use these key-points to regularize the projection matrix of the adopted low-rank transformer network (please refer to the main paper for more details). By this regularization, we hope to guide the network to use the geometry features on the extracted n key-points for a more effective prior adaptation. In Figure 5, we visualize the by-product of our SGPA, that is, n key-points of the object point cloud. Generally, the extracted n key-points can precisely localize the distinctive structure of the object (e.g., mug handle and laptop scree corner), and uniformly cover the whole region of the instance at the same time.

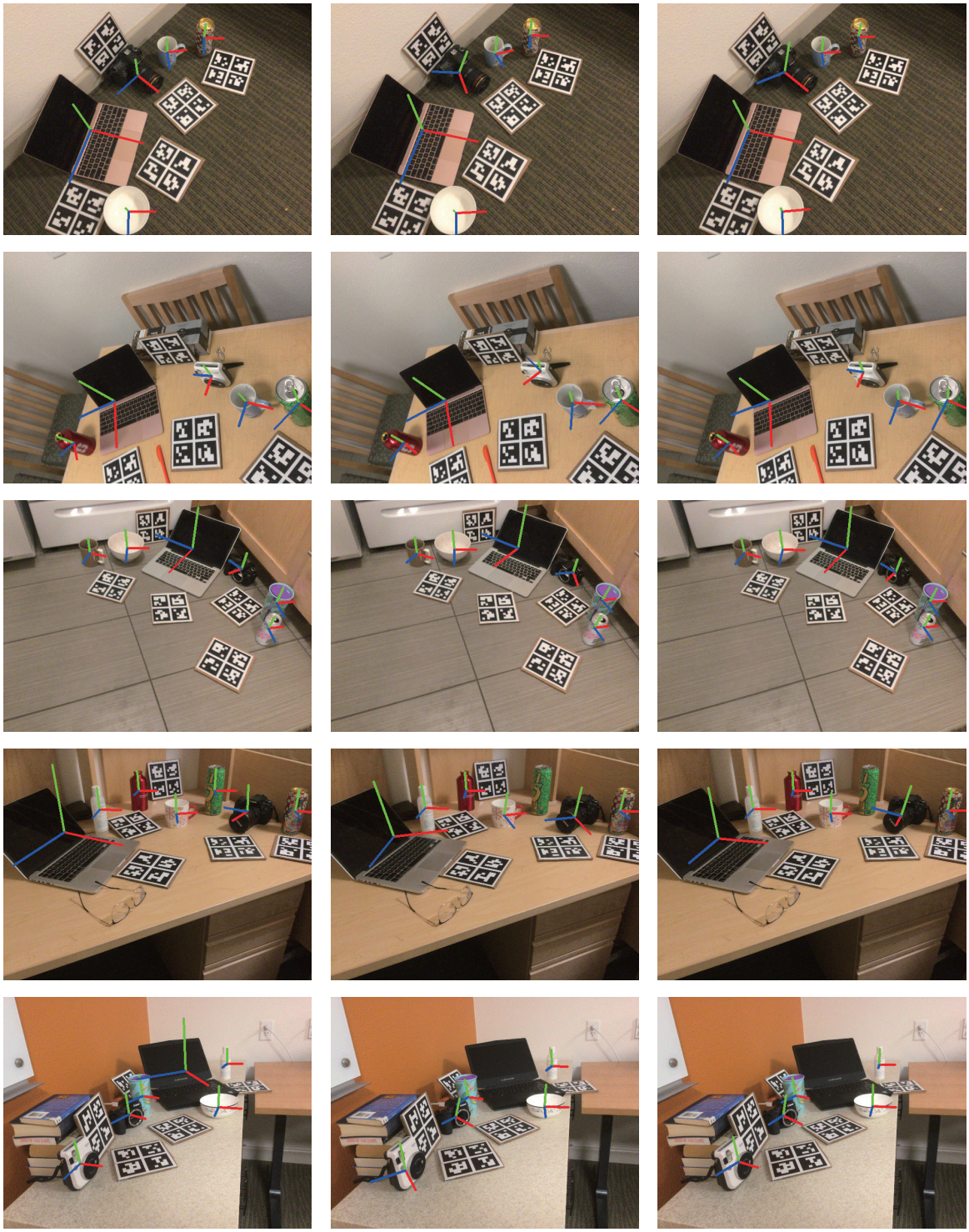
E. General Effectiveness of Prior Adaptation

In order to further investigate the effectiveness of our SGPA network, we directly deploy our prior adaptation module on SPD [1]. We trained SPD and SDP + prior adaptation with the same setting, and evaluated their pose estimation performance on both CAMERA25 and REAL275 datasets. Table 1 gives the comparative results. Compared with SPD, adding our proposed prior adaptation module significantly improves the pose accuracy on both datasets. Specifically, it improves the mAP of IoU_{75} and $5°2cm$ from 85.5% and 58.1% to 87.5% and 64.9% on the CAMERA25, and from 56.7% and 20.0% to 58.8% and 26.5% on the REAL275. These results further demonstrate the effectiveness of our proposed prior adaptation method for category-level 6D object pose estimation.

References

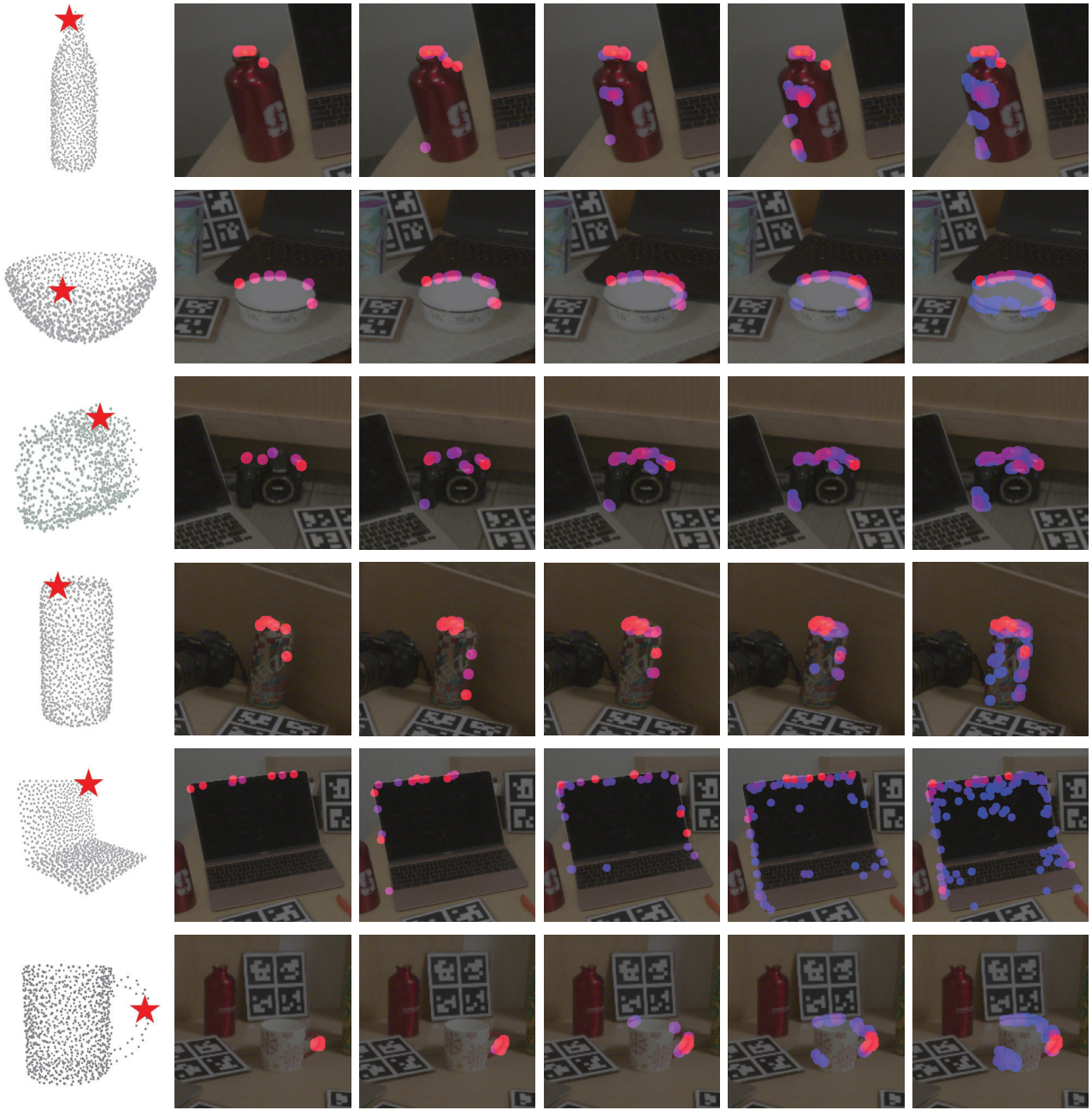
[1] Meng Tian, Marcelo H. Ang Jr, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[2] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 1



(a) Ground-truth (b) SPD (c) Ours SGPA

Figure 3. **Qualitative comparisons between our SGPA and SPD (Section B).** Three axes associated with each object indicate the pose estimation result. If the axis position and orientation are more consistent with the one of ground truth, the pose estimation accuracy of the algorithm is higher. Best viewed in color with zoom in.



(a) Query point (b) Top-8 attention (c) Top-16 attention (d) Top-32 attention (e) Top-64 attention (f) Top-128 attention

Figure 4. **Visualization of attention map (Section C).** We visualize the attention map learned by our prior adaptation network. For each query point on the prior point cloud (a), we visualize its relationships with the object point cloud by projecting the attention values onto the corresponding RGB image. To show the trend of the attention map, we present Top-k attentions, where $k = 8, 16, 32, 64, 128$ from (b) to (f). The color varies from blue to red corresponding to the attention value varies from small to large.

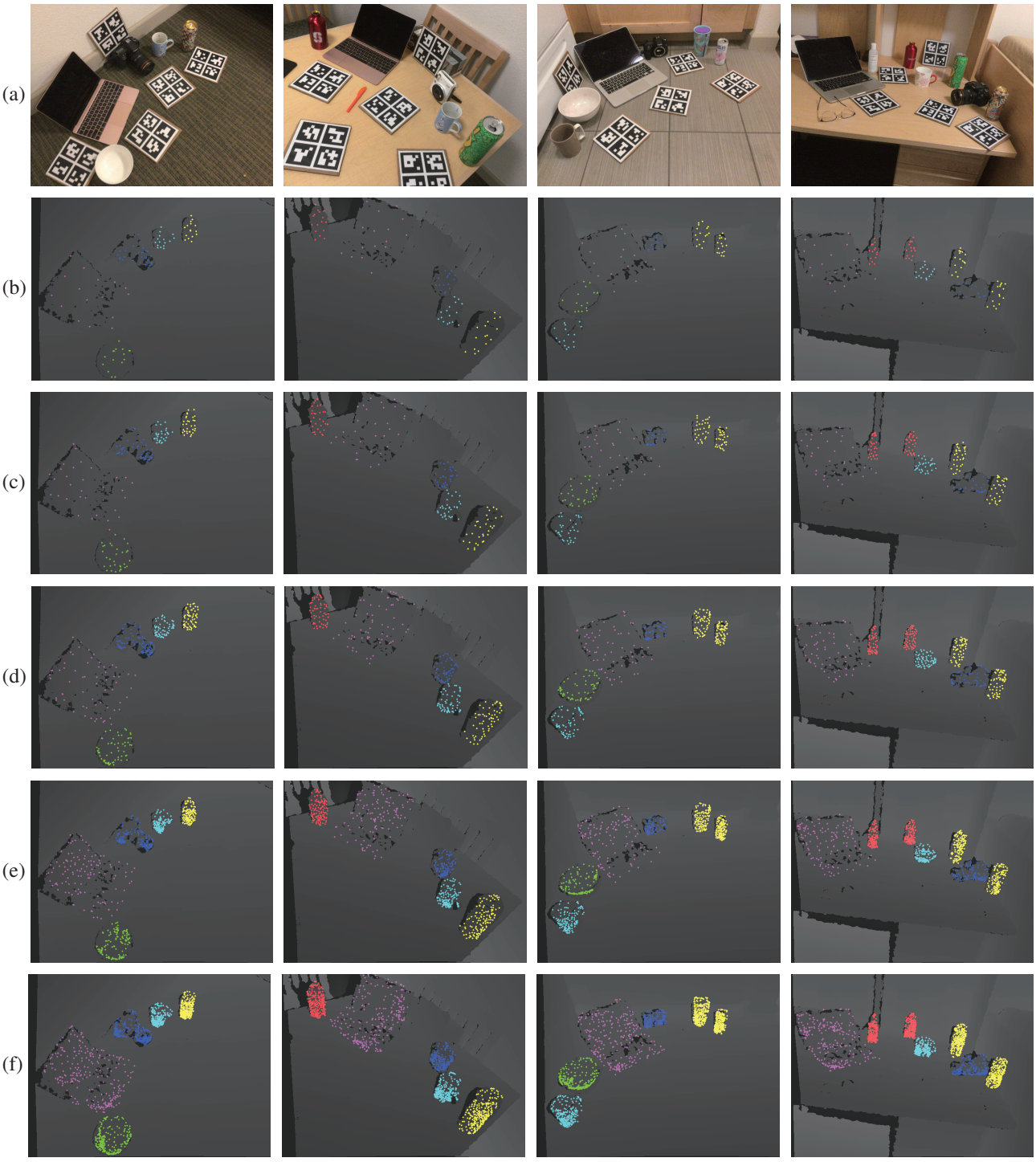


Figure 5. **Object key-point visualization (Section D)**. We visualize the sparse key-points extracted by our structure regularized low-rank transformer network. Key-points are projected onto the depth image for a clear visualization. For each column, (a) is the RGB observation. (b)-(f) are results correspond to 16, 32, 64, 128 and 256 key-points respectively. Different colors indicate key-points for different categories of objects.