Appendix: Learning to Detect Scene Landmarks for Camera Localization

Tien Do[1] Ondrej Miksik[2] Joseph DeGol[2] Hyun Soo Park[1] Sudipta N. Sinha[2]

[1] University of Minnesota [2] Microsoft

In this supplementary document, we present additional quantitative results that could not be included in the main paper. We show extensive qualitative results from our method on the INDOOR-6 dataset and also include a supplemental video. Finally, we discuss some failure cases.

1. Quantitative Results

In this section, we show the storage efficiency of our method (NBE+SLD) compared to a retrieval and matching-based method (HLoc [3]) (Section 1.1). Next, we further compare accuracy between our method and multiple baselines through a recall plot that uses a range of thresholds (Section 1.2). Finally, we report bearing errors for predicted landmarks on INDOOR-6 and 7-SCENES [6] datasets (Section 1.3).

1.1. Storage comparison

NBE+SLD requires constant storage. Figure 1 reports the storage requirements for HLoc and our method for each scene in the INDOOR-6 dataset. Our method requires 0.135 GB of storage for the SLD and NBE networks' parameters that are constant for all the scenes. This is significantly smaller than HLoc that requires 1.5GB and 1.2GB on the two larger scenes – scene1 and scene5, respectively. HLoc stores SuperPoint [1] and SuperGlue [4] networks' parameters and SuperPoint's features and VLAD [2] image descriptors for all the database images. The storage for features grows linearly with the number of database images, and that can dominate total storage on large scenes.

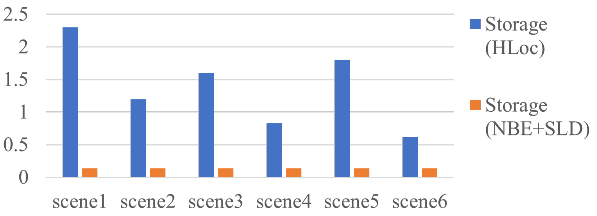


Figure 1. **Storage (in GB) used by HLoc and NBE+SLD (Ours) on INDOOR-6.** Our method uses constant storage. In comparison, HLoc requires more storage for scenes with more database images.

1.2. Recall plots on INDOOR-6 dataset

We further study the relative performance between the proposed methods NBE+SLD (E), NBE+SLD, HLoc+SLD versus DSAC*, HLoc, HLoc_L$_{300}$, HLoc_L$_{1000}$, HLoc$_{-3000}$ by analyzing the recall of the pose estimate within x degrees and x centimeters where x varies from 0 to 10. To that end, we continuously vary the recall threshold from $(0^\circ, 0\ cm)$ to $(10^\circ, 10\ cm)$ (see Figure 2). We observe that our method NBE+SLD with both EfficientNet-Lite0 and ResNet-18 backbones achieves consistently higher ac-

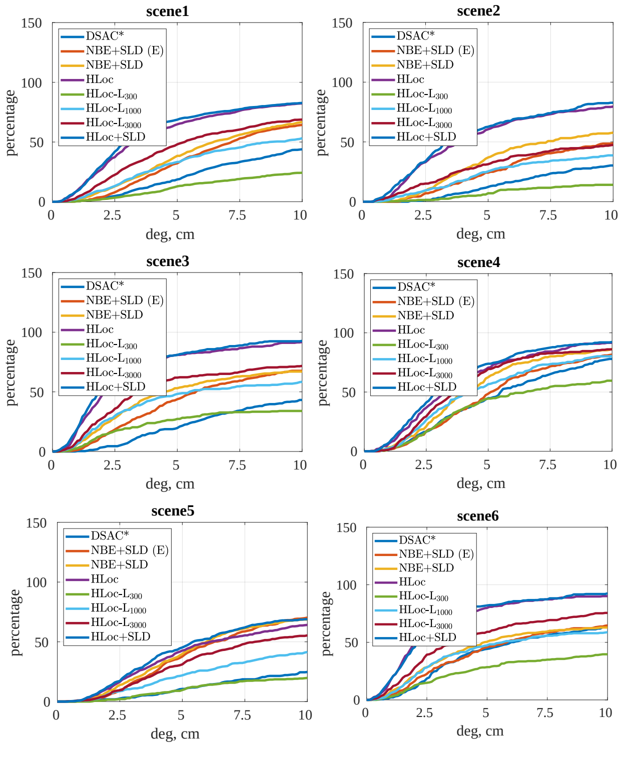


Figure 2. **Recall plots.** We show the recall of the pose estimate; *i.e.*, the percentage of test images with rotation and position error less than x degree and x centimeter, respectively, for DSAC*, NBE+SLD (Ours), HLoc, and HLoc+SLD (Ours) on each scene of our INDOOR-6 dataset.

Method	INDOOR-6											
	scene1		scene2		scene3		scene4		scene5		scene6	
	mean↓	median↓	mean↓	median↓	mean↓	median↓	mean↓	median↓	mean↓	median↓	mean↓	median↓
NBE (ours)	6.12	4.20	9.17	5.60	7.96	5.17	11.07	6.38	8.96	6.23	10.02	6.46
SLD (ours)	2.91	0.28	5.43	0.30	2.72	0.21	1.35	0.29	5.23	0.29	2.44	0.22

Table 1. **Landmark Bearing Estimation Error on INDOOR-6.** We present the angular error (in degrees) for the predicted landmark bearings from NBE and SLD on the test set of INDOOR-6. In both cases, we compute the ground truth bearing landmark vector using the ground truth camera poses and the known 3D landmark positions. For SLD, we compute the error only for the detected landmarks whose heatmap peak value is > 0.2. We observe that heatmap prediction (SLD) is always more accurate than bearing regression (NBE).

Method	7-SCENES													
	chess		fire		heads		office		pumpkin		redkitchen		stairs	
	mean↓	median↓	mean↓	median↓	mean↓	median↓	mean↓	median↓	mean↓	median↓	mean↓	median↓	mean↓	median↓
NBE (ours)	4.31	2.84	4.89	3.1	5.57	3.74	5.21	2.99	4.76	2.95	4.72	3.09	10.83	9.7
SLD (ours)	0.39	0.11	0.46	0.14	1.06	0.19	1.23	0.16	0.7	0.13	0.79	0.12	2.41	0.29

Table 2. **Landmark Bearing Estimation Error on 7-SCENES.** We present the angular error (in degrees) for the landmark predicted by NBE and SLD on the test set of 7-SCENES. In both cases, we compute the ground truth bearing landmark vector using the ground truth camera poses and the selected points' 3D locations. For SLD, we only compute the error for the detected landmarks whose heatmap peak value is > 0.2. For both NBE and SLD, we observe that the accuracy on 7-SCENES is higher than on the more challenging INDOOR-6 dataset. Results in Table 1 and 4 in the main paper also indicates that the pose accuracy is strongly correlated with bearing accuracy.

curacy than DSAC* on all scenes at all thresholds. In Figure 12, we show failure examples which are mainly due to an insufficient number of visible landmarks. However, as discussed in the main paper, the benefit of combining HLoc and SLD is clear, and HLoc+SLD outperforms HLoc by a notable margin on the scene2 and scene5.

1.3. Bearing Landmark Error for NBE and SLD

SLD is much more accurate than NBE. We present the angular error (in degrees) for the predicted bearing landmark by NBE and SLD on the test set of INDOOR-6 and 7-SCENES in Table 1 and Table 2 respectively. We compute the ground truth bearing landmark vector using the ground truth camera poses and the 3D location of the selected landmarks. There is no ground truth landmark visibility for the test set images, so, for SLD, we compute the error for the detected landmarks whose heatmap peak value is greater than 0.2. First, we note that heatmap-based SLD achieves sub-degree accuracy; e.g., median angular error ranges from 0.22 to 0.3 degrees on the INDOOR-6 and 0.11 to 0.29 degrees on the 7-SCENES.[1] In contrast, regression-based NBE while attaining 100% recall has much lower bearing accuracy. Thus, to obtain both high accuracy and high recall, one needs to combine the global (NBE) and local (SLD) predictions.

INDOOR-6 is more challenging than 7-SCENES. In both cases of NBE and SLD on the 7-SCENES dataset, we observe that the overall accuracy is significantly higher than on the INDOOR-6 dataset; e.g., the median of angular error of SLD on 7-SCENES is roughly half those in INDOOR-6, suggesting that our dataset is more challenging than the 7-SCENES dataset. In addition, by incorporating Table 1 and Table 4 in Section 4 (camera pose estimation on INDOOR-6 and 7-SCENES, respectively), we can conclude that the camera pose accuracy is highly dependent on the accuracy of the bearing landmark vectors, suggesting that one future research direction for improving camera pose estimation is to increase the accuracy and recall on the landmark bearing vector prediction.

2. Qualitative Results

In this section, we first discuss some features of our INDOOR-6 dataset (Section 2.1). We then present more results from our landmark selection approach (Section 2.2). Finally, we show extensive qualitative results for several test images from INDOOR-6, which includes visualizations for the detected landmarks and for assessing the accuracy of the estimated camera pose (Figures 6–11).

Supplementary video. In the supplementary video, we show landmarks detected by SLD on some 30 FPS video segments from the test sequences for various scenes. Specifically, we visualize the predicted heatmap values for each landmark obtained using the SLD architecture. We can observe that certain landmarks have strong and stable colors in the videos. These are the ones reliably detected on a series of subsequent frames. The presence of flickering and lighter color tones indicate that the landmark detections have higher uncertainty and are sometimes not detected.

[1] Since we cannot show the recall of the detected landmarks due to the lack of ground truth visibility on test set, we instead present the percentage of images that observe more than 8 landmarks in the main paper (see Table 2 in the main submission). This metric is strongly correlated to the recall and directly influences the camera pose estimation.

2.1. Dataset

In this section, we discuss some features of our INDOOR-6 dataset. As discussed in the main paper, our dataset contains multiple sequences from the same scene that were recorded at different times of day over a period of several days. These images contain noticeable lighting variations and were captured under indoor lighting or low-light conditions. The scenes depict real homes where the scene changes with time, and there are many examples of non-stationary objects being moved and doors and windows being opened and closed. Figure 3 shows some example images from scene1, scene4, and scene5.

We generate the pseudo ground truth camera poses for the train and test images via a two-step process. We first run COLMAP [5] only on the train images and use that reconstruction to select landmarks and derive the poses of the train images. Next, we run COLMAP on all the images in both the train and test set and obtain a second reconstruction that we robustly register to the first one by estimating a 3D similarity transformation. By employing this two-step process, we ensure that our landmarks and the models trained to detect them are derived solely from the train images, while ensuring that the camera poses for the test images are estimated quite accurately. Finally, the reconstructions are scaled to real-world dimensions using measurements of known objects.

2.2. Landmark Selection

We present the qualitative results for the selected landmarks by Algorithm 1 (see Section 3.5) in Figures 4–5. Specifically, Figure 4 shows the top-50 landmarks with highest salient scores from the training point cloud. We observe that our algorithm selects the landmark that ensures uniform coverage over the entire scene. Note that having multiple episodes per scene allows us to select highly salient landmarks that are likely to be stationary and recognizable at different times (see the color squares on Figures 6–11).

However, it is evident that with 50 landmarks, one cannot ensure a robust camera localization due to an insufficient number of visible landmarks (≥ 8) for most viewing directions. Figure 5 illustrates the effect of scene coverage when selecting more landmarks (200, 300) on scene5 and scene6. Note that for scene5, the landmarks have a pretty uniform distribution, which is desirable. Whereas, in scene6, the non-homogeneous point distribution in the original SfM point cloud and the textureless surfaces in the scene leads to a non-uniform distribution of landmarks which are clustered into multiple dense areas (especially when 300 landmarks are selected). The low coverage is partly responsible for low recall on scene6. Finally, although selecting more landmarks improves scene coverage, training the network to accurately distinguish visually similar landmarks becomes challenging. Thus, it is worth exploring new tightly coupled strategies to jointly select landmarks during the training pipeline.

References

[1] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR workshops*, 2018. 1

[2] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, 2010. 1

[3] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 1

[4] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 1

[5] Johannes Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *CVPR*, 2016. 3

[6] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *CVPR*, 2013. 1

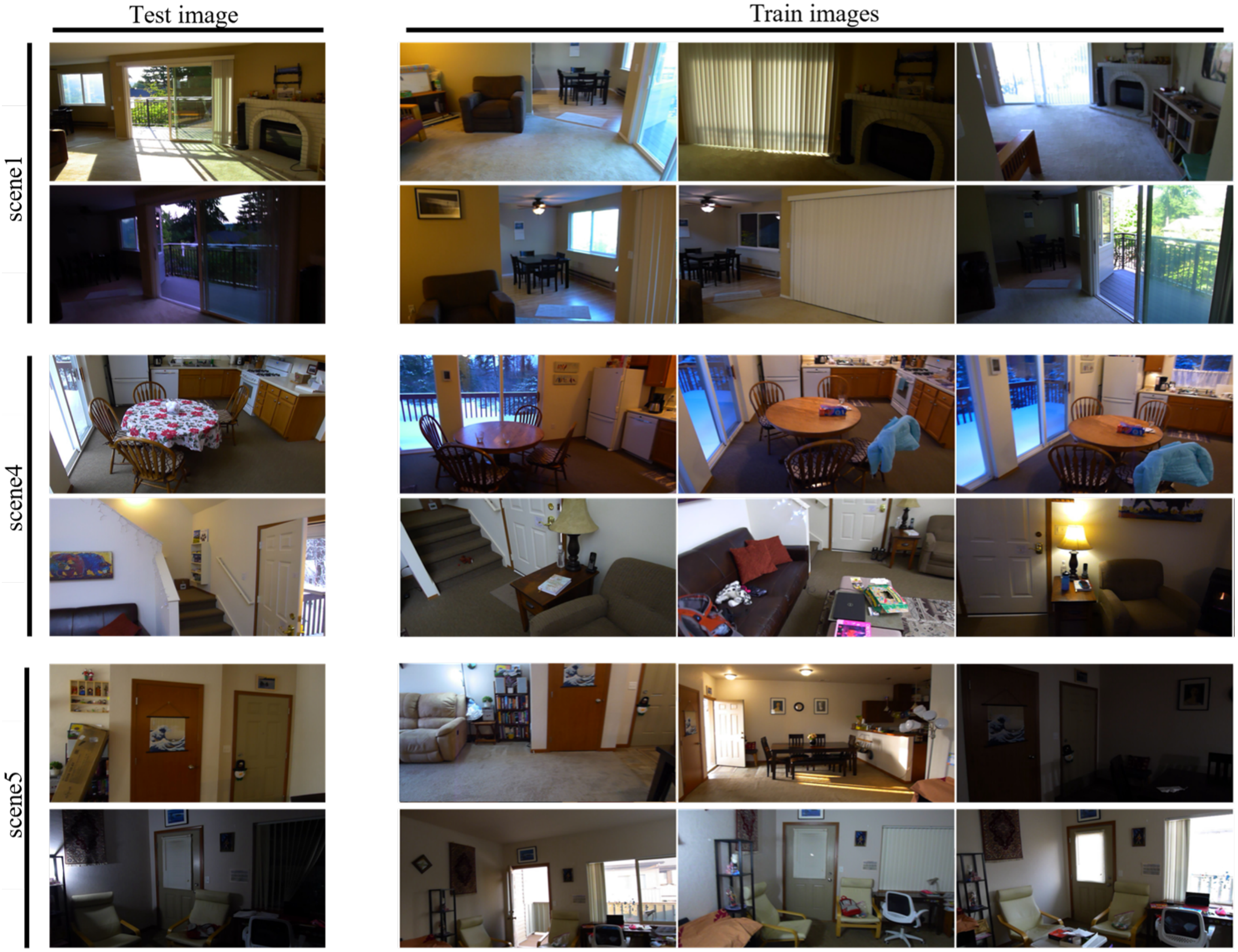
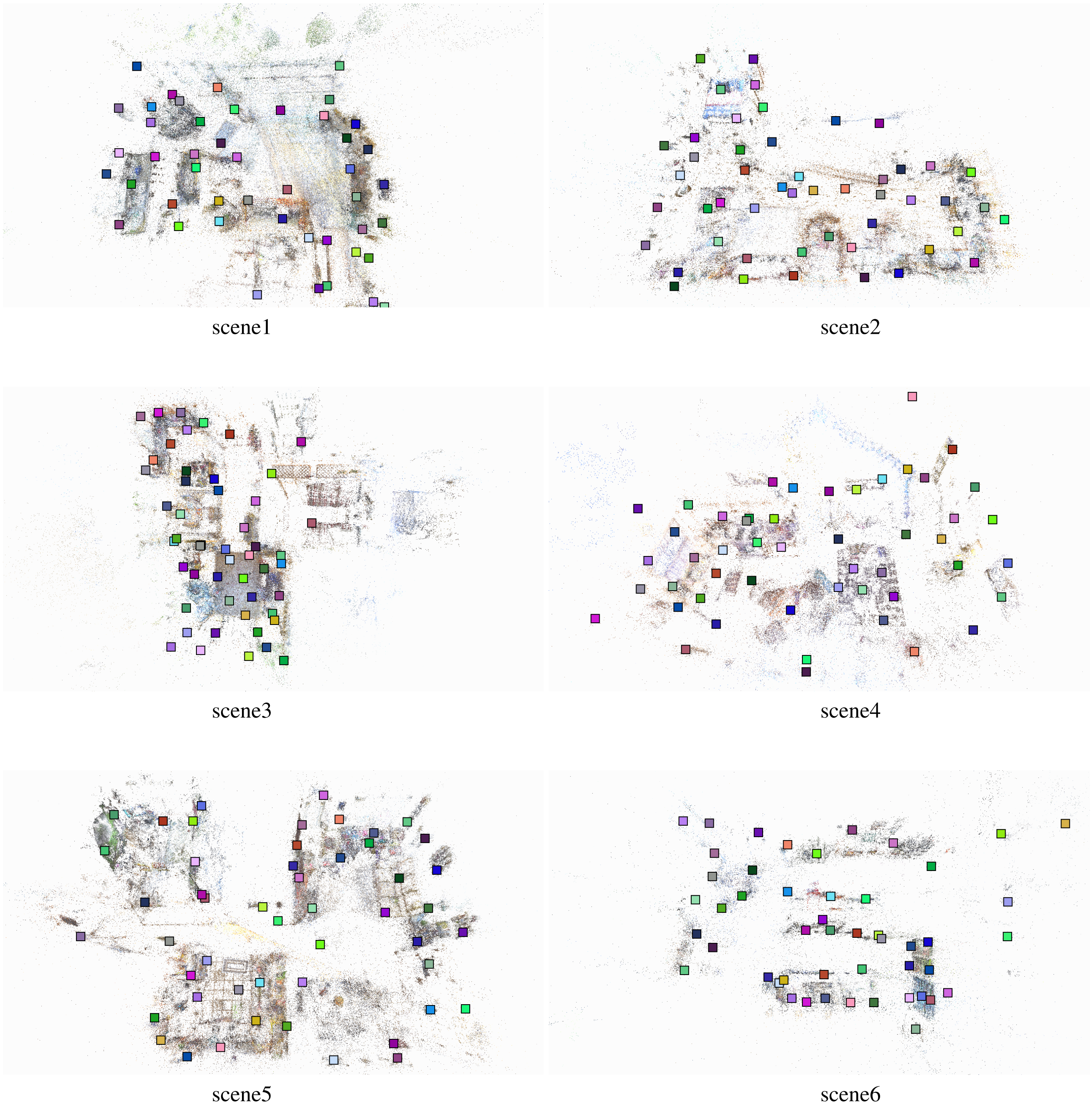


Figure 3. **INDOOR-6 DATASET:** Each row shows a test image and three overlapping training images for the scene1, scene4, and scene5. These images demonstrate the presence of notable illumination variation; e.g., induced by time-of-day, changes in scene geometry caused by furniture or objects being moved, windows and doors being opened and closed, and other forms of scene dynamics.


scene1

scene2

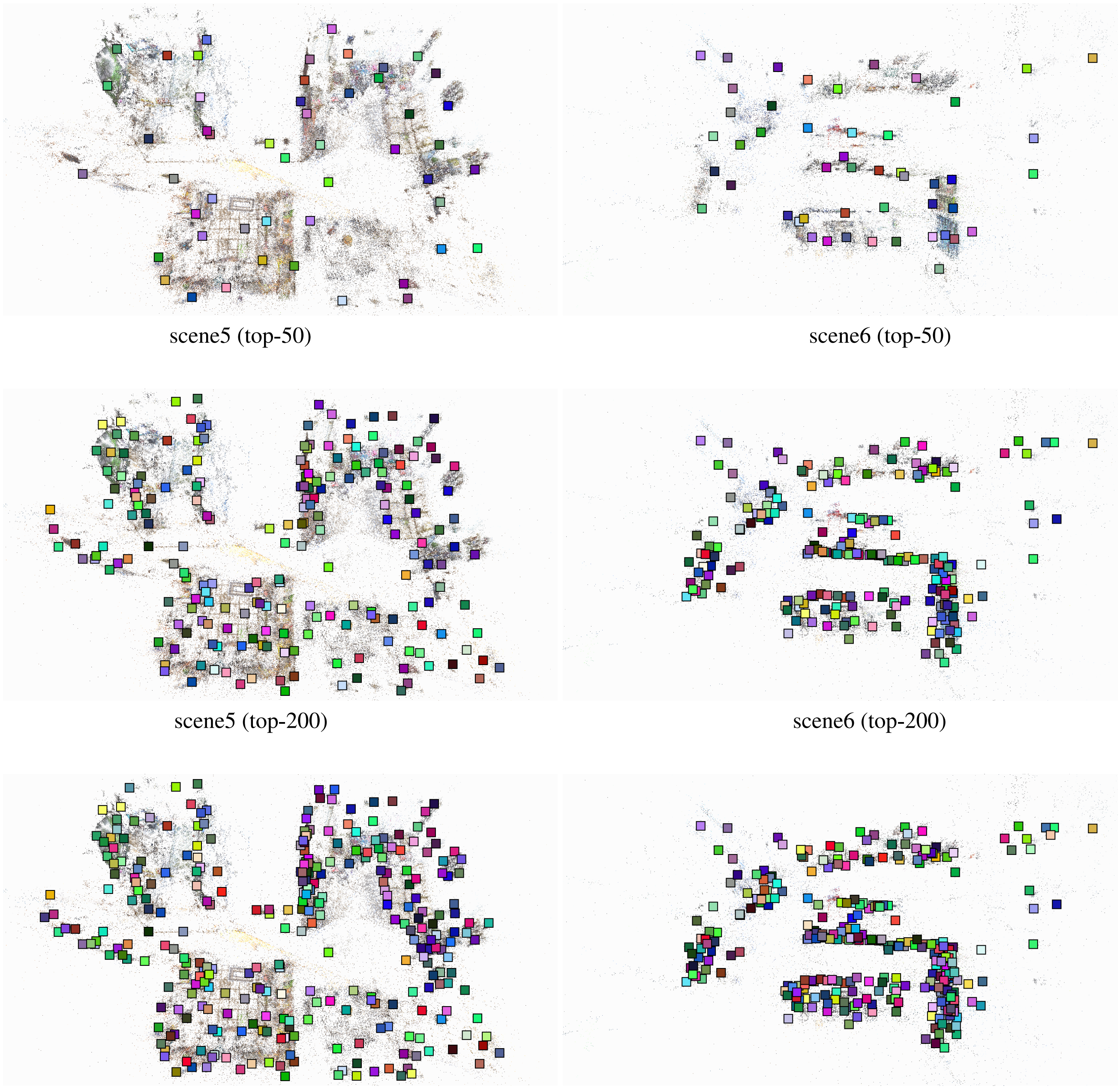
scene3

scene4

scene5

scene6

Figure 4. **Top-50 selected scene landmarks.** We visualize the 3D locations of the top-50 landmarks selected by our method for the scenes in the Indoor-6 dataset. The colored squares indicate different landmarks, and they are shown overlaid on the 3D point cloud obtained by running SfM on the training images. The selected landmarks mostly lie on scene surfaces or objects which did not move between multiple episodes in the training sequences (wall hangings, stationary furniture, etc.).



scene5 (top-50)

scene6 (top-50)

scene5 (top-200)

scene6 (top-200)

scene5 (top-300)

scene6 (top-300)

Figure 5. **Varying the number of scene landmarks.** We show the top-50, 200, and 300 landmarks selected by our method on the scene5 and scene6 scenes. Note that the top 200 here includes the top 50, and so on. Selecting more landmarks improves scene coverage and pose recall but accurately distinguishing them from one another can also become challenging.

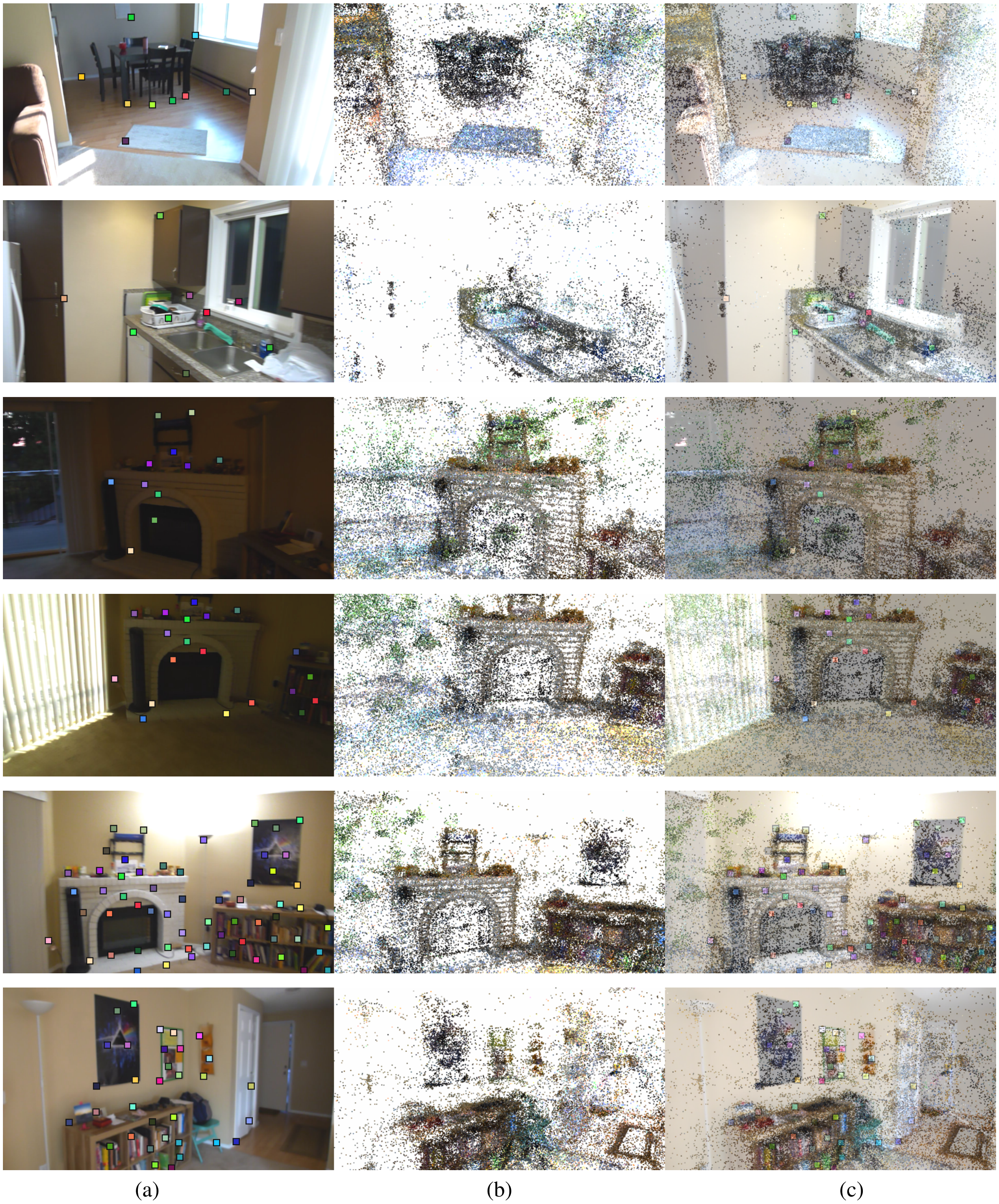


Figure 6. **Qualitative results (scene1).** Each row shows (a) a query image from the test set with SLD landmark detections (shown as colored squares); (b) the rendering of the SfM point cloud obtained from training images, projected using the camera pose estimated by our proposed NBE+SLD method; and (c) an overlay of the rendering on the query image. The point cloud is not used by our method. They are rendered only for the purpose of visualization to assess the accuracy of the pose estimate. The images contain dramatic illumination changes in this scene, ranging from strong sunlight to low-light at night. Our method detects several landmarks despite the lighting variations.

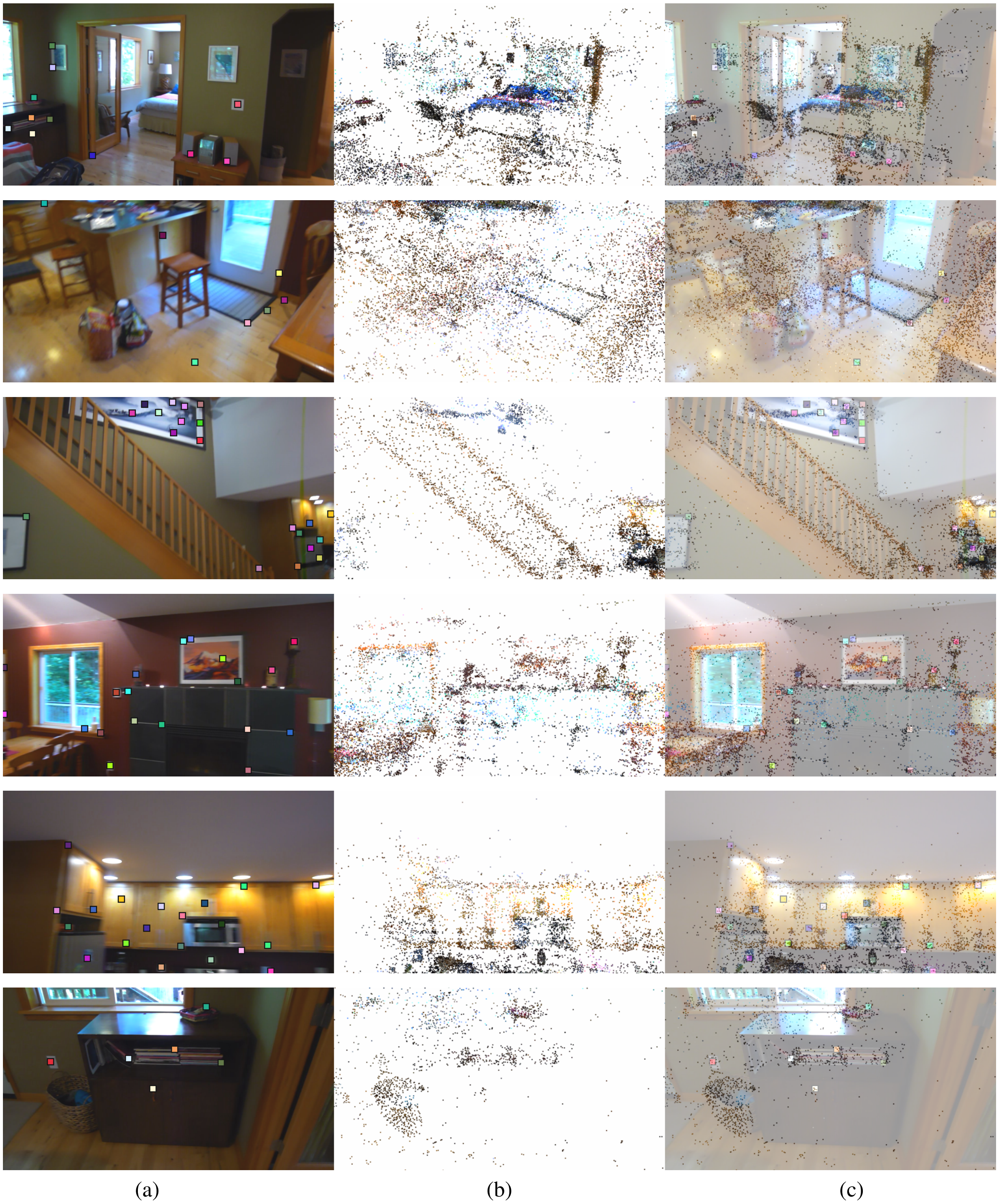


Figure 7. **Qualitative results (scene2):** Each row shows (a) a query image from the test set with SLD landmark detections (shown as colored squares); (b) the rendering of the SfM point cloud obtained from training images, projected using the camera pose estimated by our proposed NBE+SLD method; and (c) an overlay of the rendering on the query image. The point cloud is not used by our method. They are rendered only for the purpose of visualization to assess the accuracy of the pose estimate. On this scene, often too few landmarks (≤ 8) are detected by SLD. Thus, the pose is often estimated using both SLD detections and NBE bearing predictions.

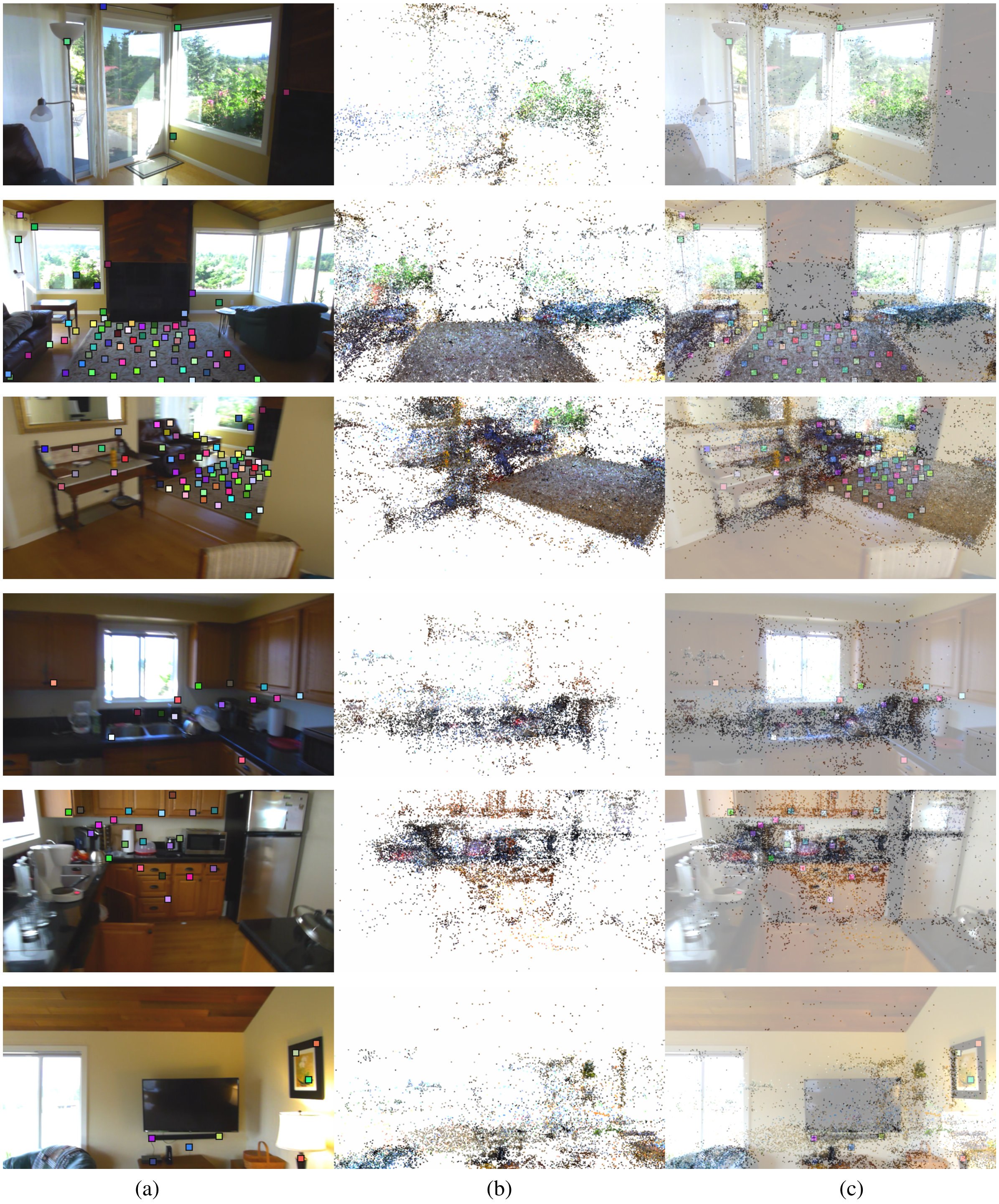


Figure 8. **Qualitative results (scene3):** Each row shows (a) a query image from the test set with SLD landmark detections (shown as colored squares); (b) the rendering of the SfM point cloud obtained from training images, projected using the camera pose estimated by our proposed NBE+SLD method; and (c) an overlay of the rendering on the query image. The point cloud is not used by our method. They are rendered only for the purpose of visualization to assess the accuracy of the pose estimate.

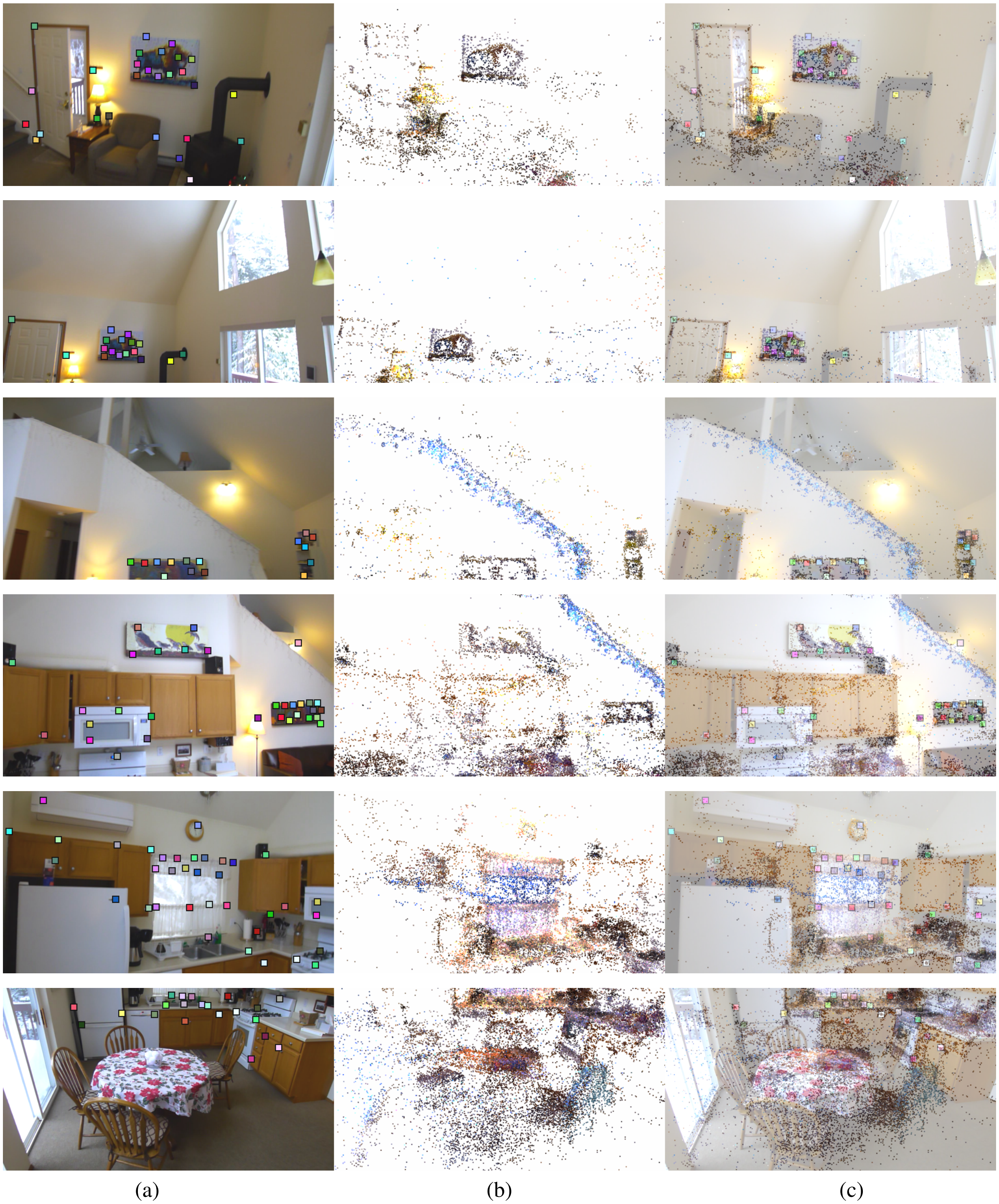


Figure 9. **Qualitative results (scene4):** Each row shows (a) a query image from the test set with SLD landmark detections (shown as colored squares); (b) the rendering of the SfM point cloud obtained from training images, projected using the camera pose estimated by our proposed NBE+SLD method; and (c) an overlay of the rendering on the query image. The point cloud is not used by our method. They are rendered only for the purpose of visualization to assess the accuracy of the pose estimate. The last row shows an example with noticeable scene change: the table cloth wasn't present earlier and the chairs were in different positions in the training sequences. Our method, however, uses persistent landmarks in the background to compute an accurate pose.

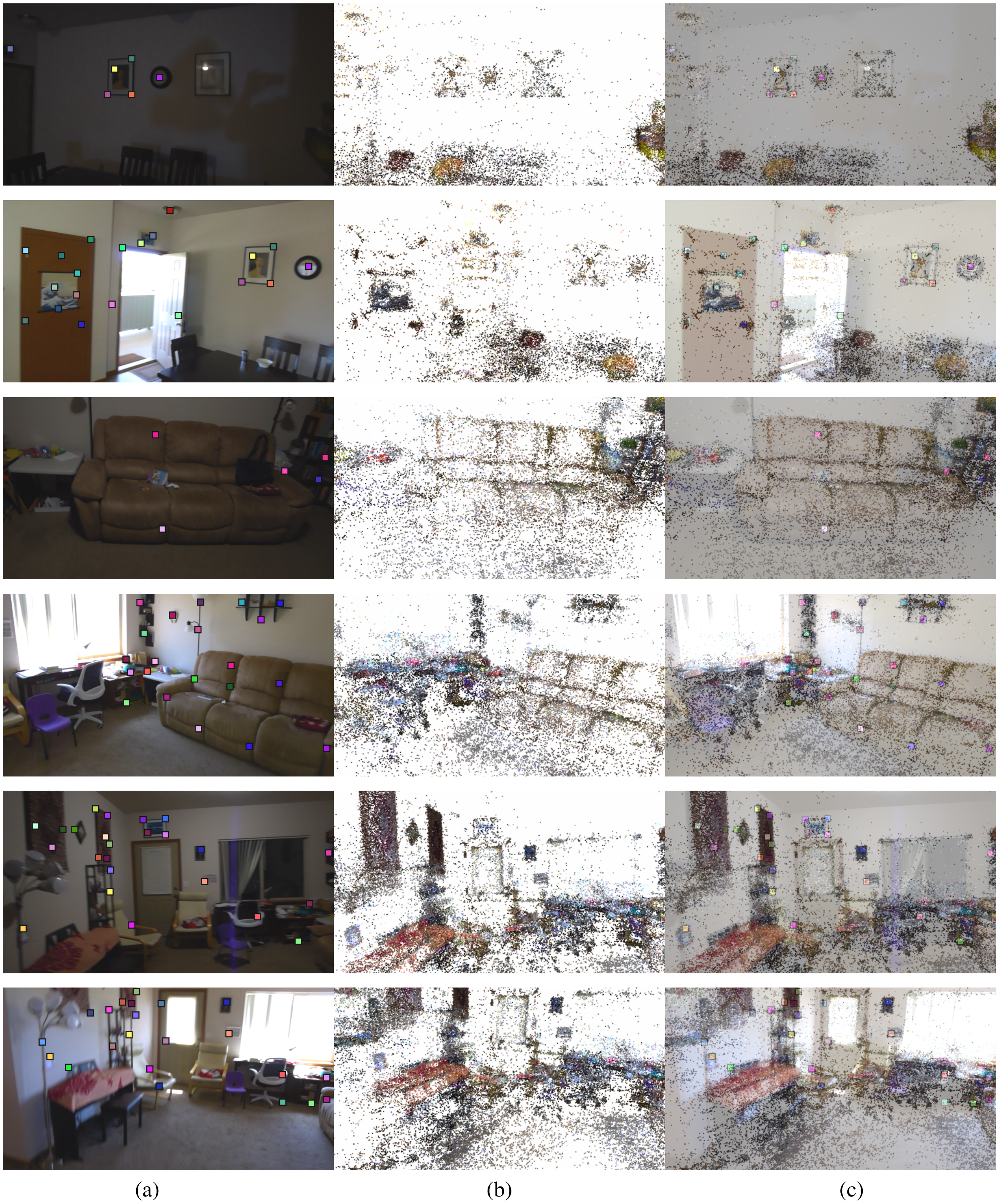


Figure 10. **Qualitative results (scene5):** Each row shows (a) a query image from the test set with SLD landmark detections (shown as colored squares); (b) the rendering of the SfM point cloud obtained from training images, projected using the camera pose estimated by our proposed NBE+SLD method; and (c) an overlay of the rendering on the query image. The point cloud is not used by our method. They are rendered only for the purpose of visualization to assess the accuracy of the pose estimate. These images show SLD detections on the same objects and surfaces (photographs, wall hangings, sofa, door corner) in different images where the lighting varies considerably.

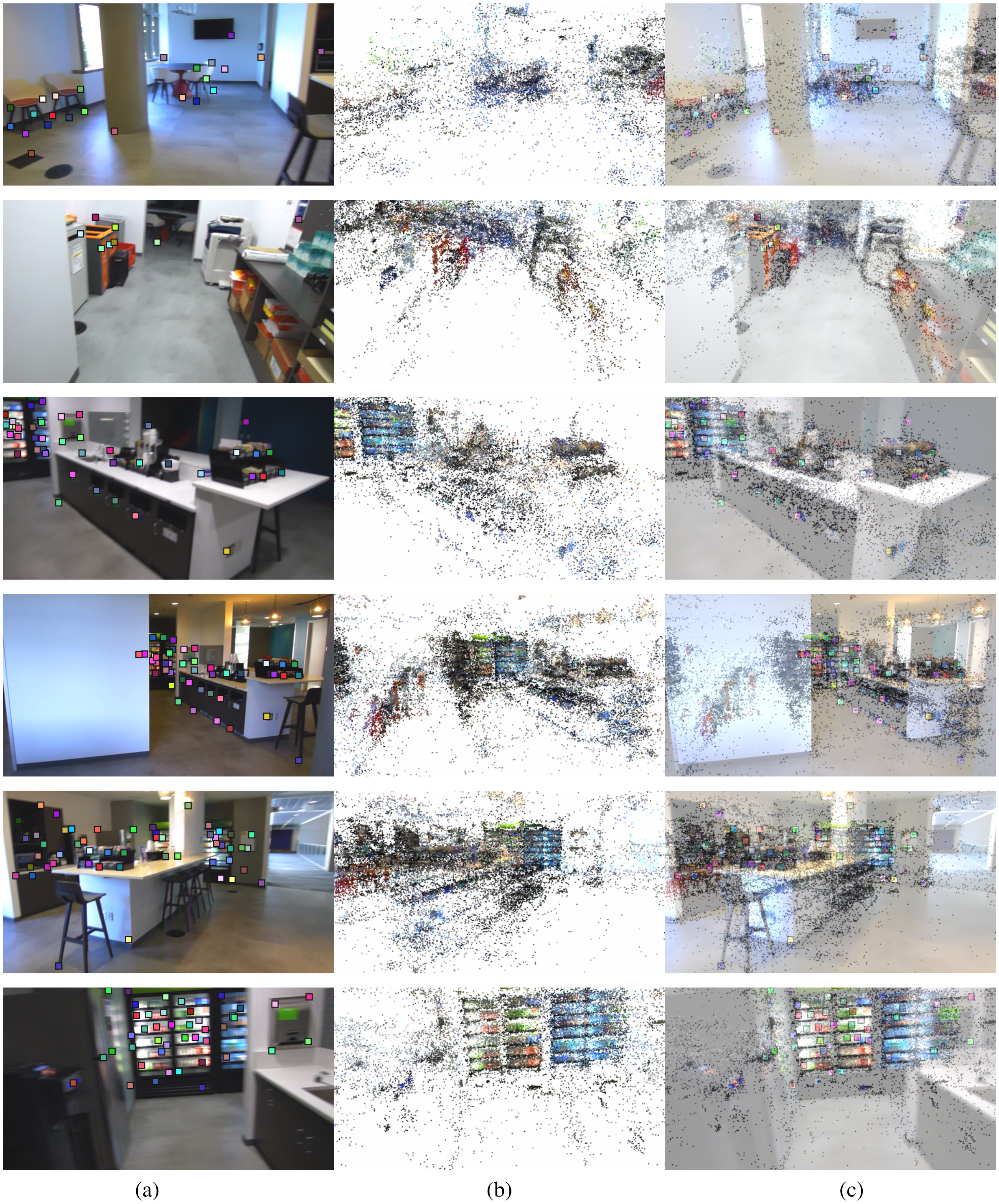


Figure 11. **Qualitative results (scene6):** Each row shows (a) a query image from the test set with SLD landmark detections (shown as colored squares); (b) the rendering of the SfM point cloud obtained from training images, projected using the camera pose estimated by our proposed NBE+SLD method; and (c) an overlay of the rendering on the query image. The point cloud is not used by our method. The point cloud rendering is used purely for visualization to confirm the accuracy of the pose estimate.

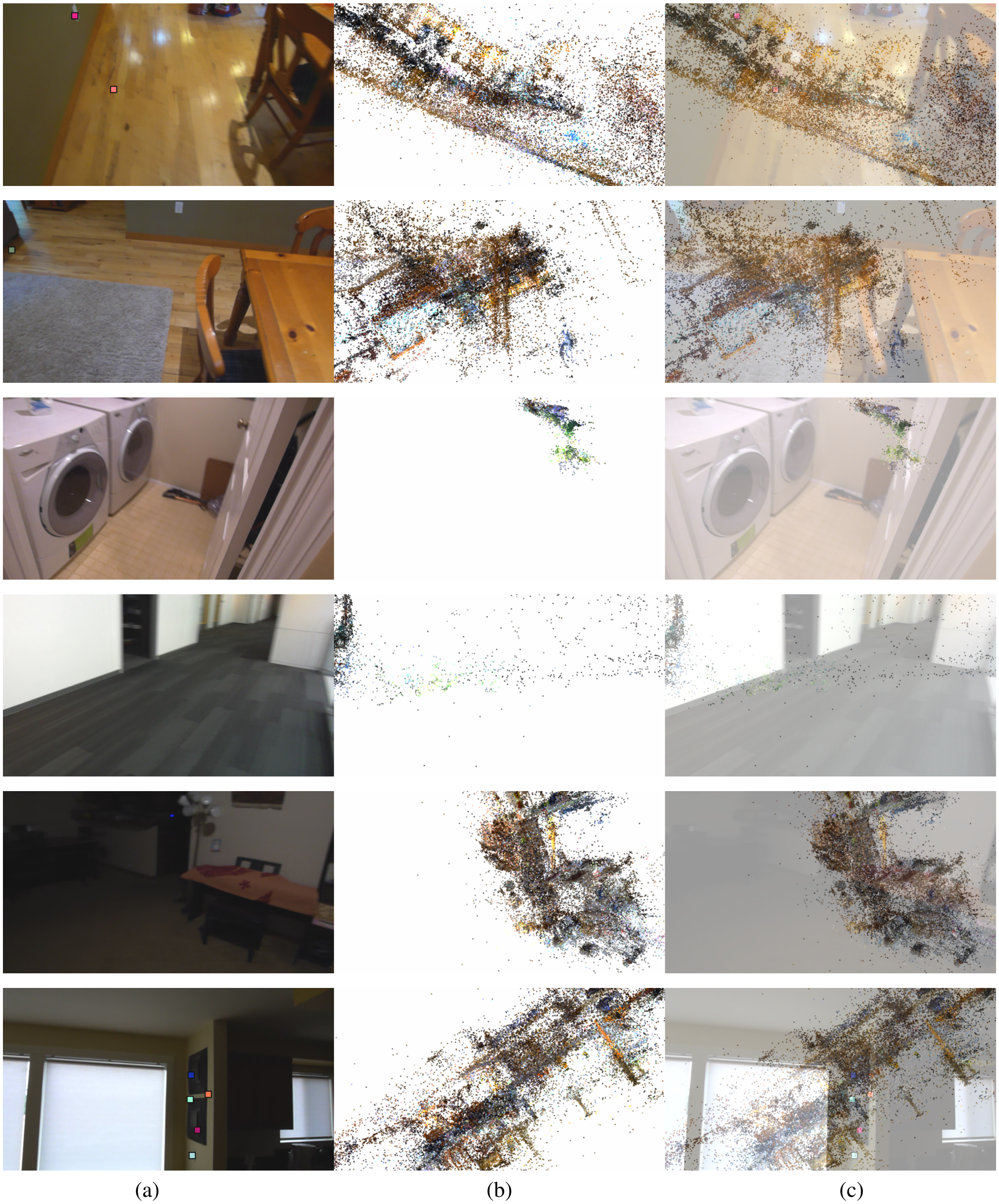


Figure 12. **Failure Examples:** Each row shows (a) a query image from the test set with SLD landmark detections (shown as colored squares); (b) the rendering of the SfM point cloud obtained from training images, projected using the camera pose estimated by our proposed NBE+SLD method; and (c) an overlay of the rendering on the query image. The point cloud is not used by our method. Our method can be inaccurate when a sufficient number of landmarks are not detected and NBE bearing predictions are also not accurate enough. Some failures are shown (top to bottom) – insufficient landmarks and too few training images observing floor (scene2), laundry room (scene1), and hallway (scene6); too dark (scene5); and high dynamic range-induced false detections (scene3).