Supplementary Material for

CPO: Change Robust Panorama to Point Cloud Localization

Junho Kim[1], Hojun Jang[1], Changwoon Choi[1], and Young Min Kim[1,2]

[1] Department of Electrical and Computer Engineering, Seoul National University
[2] Interdisciplinary Program in Artificial Intelligence and INMC, Seoul National University

A Structured3D Dataset Details

In this section, we provide the details for preparing the Structured3D [14] dataset. Due to copyright constraints, the 3D models of the dataset is unavailable to the public. Therefore we generated synthetic 3D meshes using the layout annotations and color values from the panorama images, where qualitative samples are shown in Figure A.1. As explained in Section 4.1, Structured3D contains 21845 rooms from which we select 672 rooms for evaluation, and each room has three object configurations (empty, simple, full). We create the 3D model using the empty object layout and set the query panorama as the full object layout. To additionally evaluate illumination robustness, we randomly choose the lighting setup from the three possible configurations (raw, cold, warm) for each object configuration.

B Baseline Details

In this section, we describe the details for implementing the baselines compared against CPO. As we implement PICCOLO [10] using the publically available codebase released by the authors, we focus our description on the Structure-based and depth-based approaches. For fair comparison, we set the translation/rotation starting points N_t, N_r and the number of candidate poses K identical to CPO.

Structure-Based Approach As explained in Section 4, structured-based approach first finds promising candidate poses using robust image retrieval and then refines poses using PnP-RANSAC from feature matches. For image retrieval we use OpenIBL [8], which is a widely used image retrieval method that outputs a global feature vector for each image. To deploy OpenIBL in out setup, we first render $N_t \times N_r$ synthetic views from the point cloud. Then, we extract the global features for each synthetic view and the query image, and choose the top K synthetic views whose feature vectors are closest to that of the query image. As the final step, we perform feature matching [13] from each chosen synthetic view

Fig. A.1. Visualization of synthesized 3D point clouds in Structured3D [14].

against the query image, and determine the final view with the most matches. The pose from the final view is refined with feature matches from the previous step via PnP-RANSAC [6].

Depth-Based Approach Inspired from Jenkins et al. [9], depth-based approach first finds candidate poses by comparing estimated monocular depth with the 3D point cloud and refining pose with PnP-RANSAC. For monocular depth estimation we use the pretrained model from Albanis et al. [2], which can reliably estimate the underlying 3D structure from the query panorama. Then, we find the top K poses from a pool of $N_t \times N_r$ starting points that have the smallest Chamfer distance with the 3D point cloud. Similar to the structure-based approach, we perform feature matching and refine the view with the most matches via PnP-RANSAC.

C Additional Details on Score Maps

We provide additional details about score map generation. Recall that we generate 2D, 3D score maps using color consistency from histograms of synthetic views $\mathcal{Y}$. Here we generate $N_t^{\text{score}} \times N_r^{\text{score}}$ synthetic views, similar to the candidate pose selection introduced in Section 3.3. The exact number of synthetic views used to generate score maps is further specified in Section D.

D Hyperparameter Setup

In this section, we report the hyperparameter setups of CPO. As explained in Section 3.3, from $N_t \times N_r$ poses we select the top K candidate poses with the highest histogram intersection (Equation 4) for pose refinement. We follow the identical hyperparameter setup as PICCOLO [10] for pose refinement. Below we specify other hyperparameter setups that differ by the localization scenario.

D.1 Localization with Raw Color

For OmniScenes [10], Stanford 2D-3D-S [3] and Structured3D [14], where localization was done with raw color inputs, we set $N_t = 100, K = 6$. We set the number of rotation starting points as $N_r = 216$ for OmniScenes and Stanford 2D-3D-S, whereas for Structured3D we use $N_r = 24$ to run the baselines in a

reasonable amount of time. For pose selection we split the input image into 8×16 patches and generate color histograms for each patch using the fast histogram generation presented in Section 3.2. Other hyperparameter setups slightly differ by dataset, which we elaborate below.

OmniScenes Dataset As OmniScenes is mainly an indoor dataset, we employ octree-based translation starting point selection. For generating 2D and 3D score maps, we use synthetic views from $N_t^{\text{score}} = 100, N_r^{\text{score}} = 216$ poses and divide the input image into 16×32 patches. We use patches of finer scale and generate more accurate score maps to cope with large scene changes in OmniScenes [10].

Stanford 2D-3D-S Similar to OmniScenes, we employ octree-based translation starting point selection, as Stanford 2D-3D-S dataset is also an indoor dataset. For generating 2D and 3D score maps, we use synthetic views from $N_t^{\text{score}} = 100$, $N_r^{\text{score}} = 216$ poses and divide the input image into 8×16 patches.

Structured3D As explained in Section A, the 3D models in the Structured3D dataset are synthetically generated cuboids lacking clutter. Therefore we use a uniform grid partition for this dataset. Similar to the Stanford 2D-3D-S dataset, for score map generation we use synthetic views from $N_t^{\text{score}} = 100, N_r^{\text{score}} = 24$ poses and divide the input image into 8×16 patches.

D.2 Localization with Semantic Labels

For Stanford 2D-3D-S [3] and Data61/2D3D [12], where localization was done with semantic labels, we set $N_r = 216$, similar to localization with raw color. The number of translation starting points N_t differ by dataset, which is further specified below. In addition, we do not apply score maps in these scenarios as there are no scene changes in both datasets and the color values of semantic labels do not reflect any photometric information.

Stanford 2D-3D-S We employ octree-based translation starting point selection and set the number of translation starting points to $N_t = 100$, as in raw color localization. Further, we divide the input image into 8×16 patches for histogram-based initialization.

Data61/2D3D We employ grid-based translation starting point selection and set the number of translation starting points to $N_t = 300$, as the dataset is captured outdoor. Further, we confine the translation domain to a cuboid spanning $50 \times 10 \times 5m$, similar to the initialization procedure used in Campbell *et al.* [4]. The cuboid is placed to cover two lanes within the outdoor scene, which reflects the prior knowledge that the camera was mounted on a vehicle. For histogram-based initialization, we divide the input image into 4×8 patches.

(a) Original Image (b) Average Intensity (c) Gamma (d) White Balance

Fig. F.1. Synthetic color variations for evaluating illumination robustness.

Table F.1. Ablation study on color preprocessing evaluated in a subset of Stanford 2D-3D-S [3]. The images are modified by average intensity (Int.), gamma (Gam.), and white balance (W.B.).

Method	t-error (m)				R-error (°)				Accuracy			
	Orig.	Int.	Gam.	W.B.	Orig.	Int.	Gam.	W.B.	Orig.	Int.	Gam.	W.B.
CPO w/o Preprocessing	-	3.85	3.48	3.40	-	153.92	136.96	129.05	-	0.00	0.00	0.03
CPO	0.01	**0.01**	**0.01**	**0.01**	0.19	**0.21**	**0.25**	**0.25**	0.94	**0.88**	**0.88**	**0.88**

E Distortion Handling in Histogram Intersection

In this section we describe the distortion handling operation used for calculating histogram intersections in Equation 4. Since panorama images have spherical distortion, we compensate for such irregularities by applying additional weights proportional to the sin value of the latitude. To elaborate, we add an additional weight to the histogram intersection equation,

$$w(Y) = \frac{1}{2}\sum_i (M_i + S_i)\Lambda(h_i(Y), h_i(I_Q)), \tag{1}$$

where S_i is the sine value of the i^{th} patch centroid's latitude. The modified intersection equation can correctly place lesser weight on patches near the pole, as these areas are unevenly stretched in the panorama images.

F Additional Ablation Study

Color Preprocessing for Illumination Robustness We report the impact of preprocessing the color values of the panorama and point cloud for robustness against illumination changes. Recall that we match the color distributions of 2D and 3D via optimal transport, as mentioned in Section 3.1. We apply synthetic color variations to the subset of images in Area 3 from Stanford 2D-3D-S [3], as shown in Figure F.1. These images are originally used for obtaining results in Table 4 to make comparisons between CPO, PICCOLO, and GOSMA [5].

We consider three synthetic color variations: average intensity, gamma, and white balance change. For average intensity change we lower each pixel intensity by 33%. For gamma change, we set the image gamma to 3. For white balance

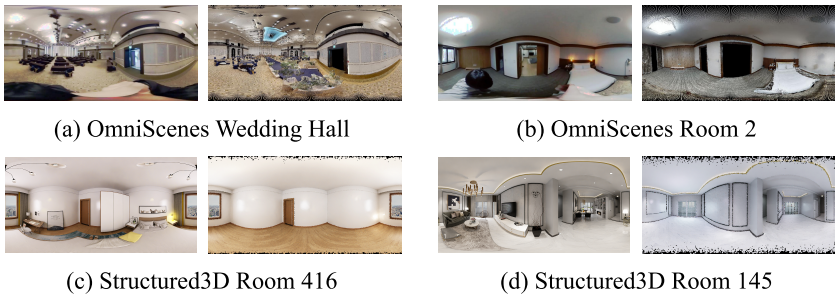

Fig. G.1. Qualitative results of CPO on OmniScenes [10] and Structured3D [14]. We display the input query image (left) and the projected point cloud under the estimated camera pose (right).

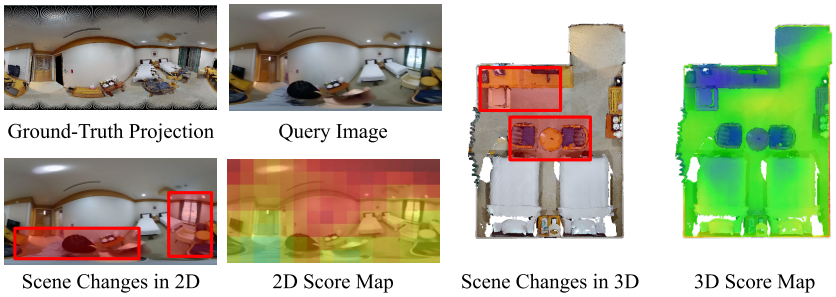

Fig. G.2. Visualization of 2D, 3D score maps. The 2D score map assigns lower scores to the capturer's hand and dislocated objects. Similarly, the 3D score map assigns lower scores to dislocated chairs and tables.

change, we apply the following transformation matrix to the raw RGB color values: $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{pmatrix}$.

Table F.1 shows the results for illumination robustness. CPO using color preprocessing shows robust performance amidst the three variations, whereas CPO without color distribution matching leads to poor performance in illumination changes. While more sophisticated color modification methods [1, 7, 11, 15] may account for complex illumination shifts, we find that our simple matching scheme suffices for handling modest color variations in practical settings.

G Additional Qualitative Results

Localization in Scenes with Changes We further report additional qualitative results of CPO in OmniScenes [10] and Structured3D [14]. As shown in Figure G.1, CPO performs robust localization under various scenes in both datasets containing large amounts of scene change.

2D, 3D Score Maps We display additional 2D and 3D score maps generated for room 4 from OmniScenes [10]. As shown in Figure G.2, the object arrangements have changed since the 3D scan. Both 2D and 3D score maps assign smaller scores to dislocated objects and the 2D score map further attenuates capturer's hand, which is not present in the 3D scan. The score maps effectively place smaller weight on regions with scene changes, leading to robust localization in CPO as demonstrated in Section 4.

References

1. Afifi, M., Barron, J.T., LeGendre, C., Tsai, Y.T., Bleibel, F.: Cross-camera convolutional color constancy. In: The IEEE International Conference on Computer Vision (ICCV) (2021)
2. Albanis, G., Zioulis, N., Drakoulis, P., Gkitsas, V., Sterzentsenko, V., Alvarez, F., Zarpalas, D., Daras, P.: Pano3d: A holistic benchmark and a solid baseline for 360° depth estimation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3722–3732 (2021). https://doi.org/10.1109/CVPRW53098.2021.00413
3. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 (2017)
4. Campbell, D., Petersson, L., Kneip, L., Li, H.: Globally-optimal inlier set maximisation for camera pose and correspondence estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence p. preprint (June 2018). https://doi.org/10.1109/TPAMI.2018.2848650
5. Campbell, D., Petersson, L., Kneip, L., Li, H., Gould, S.: The alignment of the spheres: Globally-optimal spherical mixture alignment for camera pose estimation. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. to appear. IEEE, Long Beach, USA (June 2019)
6. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981), http://dblp.uni-trier.de/db/journals/cacm/cacm24.htmlFischlerB81
7. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
8. Ge, Y., Wang, H., Zhu, F., Zhao, R., Li, H.: Self-supervising fine-grained region similarities for large-scale image localization. In: European Conference on Computer Vision (2020)
9. Howard-Jenkins, H., Ruiz-Sarmiento, J.R., Prisacariu, V.A.: Lalaloc: Latent layout localisation in dynamic, unvisited environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10107–10116 (October 2021)
10. Kim, J., Choi, C., Jang, H., Kim, Y.M.: Piccolo: Point cloud-centric omnidirectional localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3313–3323 (October 2021)
11. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. arXiv preprint arXiv:1703.07511 (2017)

12. Namin, S., Najafi, M., Salzmann, M., Petersson, L.: A multi-modal graphical model for scene analysis. In: 2015 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1006–1013. IEEE Computer Society, Los Alamitos, CA, USA (jan 2015). https://doi.org/10.1109/WACV.2015.139, https://doi.ieeecomputersociety.org/10.1109/WACV.2015.139
13. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: CVPR (2020)
14. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. In: Proceedings of The European Conference on Computer Vision (ECCV) (2020)
15. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on (2017)