Is This the Right Place?
Geometric-Semantic Pose Verification for Indoor Visual Localization
Supplementary material

Hajime Taira[1] Ignacio Rocco[2] Jiri Sedlar[3] Masatoshi Okutomi[1] Josef Sivic[2,3]
Tomas Pajdla[3] Torsten Sattler[4] Akihiko Torii[1]
[1]Tokyo Institute of Technology, [2]Inria,* [3]CIIRC CTU in Prague,† [4]Chalmers University of Technology

This supplementary material provides additional details that could not be included in the paper due to space constraints: Sec. A describes the construction of the image-scan graph in more detail (*c.f*. Sec. 3.2 in the paper). Sec. B shows that avoiding reduction of the field-of-view of a camera before extracting surface normals improves performance. Sec. C provides details on the construction of the "super-classes" (*c.f*. Sec. 3.3 in the paper) and justifies the design choice made in the paper. Sec. D details the construction of the training sets used by our trainable verification approach (*c.f*. Sec. 4 in the paper). Finally, Sec. E shows qualitative results (*c.f*. Fig. 4 in the paper).

A. Image-Scan Graph

The original InLoc dataset includes RGB-D panoramic scans and perspective RGB-D image cutouts from the scans as the database. To render more complete synthetic query images with fewer missing pixels, we construct an image-scan graph that enables us to render the synthetic images using the 3D points visible in multiple adjacent panoramic scans (*c.f*. Sec. 3.2 in the paper). Fig. A shows how we generate the graph: For each perspective database image, we compute the visual overlap with adjacent panoramic scans by projecting their 3D point clouds into the perspective database image, while taking occlusions into account. Based on the ratio of pixels in the rendered view that correspond to 3D points in the scans, we establish edges between the perspective database image and the panoramic scans that have more than 10% overlap.

B. Cropping before Normal Estimation

As mentioned in Sec. 3.2 of the paper, the original Taskonomy pipeline uses images of size 256×256 pixels as input when estimating surface normals. Using the original pipeline thus requires to crop and re-scale the images to 256×256 pixels. Since the cropping reduces the field-of-view and thus potentially discards useful information, we modified the pipeline to avoid cropping. Tab. A compares several of our pose verification methods that use normals (DenseNV and DensePNV) with and without cropping. As a reference, we also report results for DensePV [2], which does not use normal information. Using cropping reduces the performance in most cases, especially when only using normal information for verification (DenseNV). The results

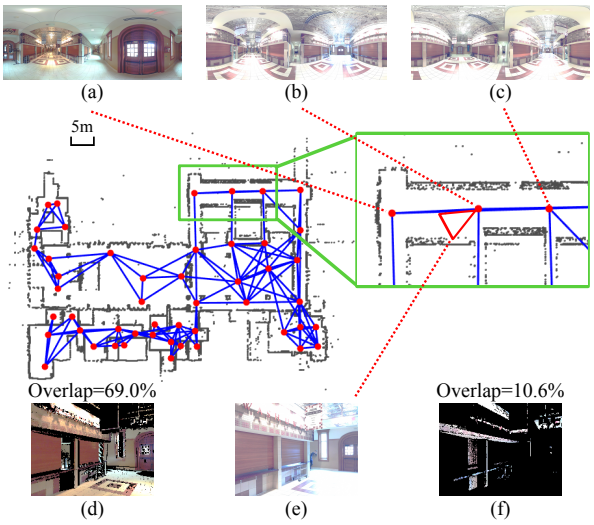


Figure A: **Image-scan graph for the InLoc dataset [2].** For each perspective database image (e) which is cut out from the RGB-D panoramic scan (b), we compute the overlap with each adjacent scan (a, c) by projecting their 3D points into the perspective view (d, f). Red dots show where RGB-D panoramic scans (and corresponding perspective database images) are located. Blue lines indicate edges between panoramic scans and perspective database images, established based on visual overlap.

*WILLOW project, Departement d'Informatique de l'École Normale Supérieure, ENS/INRIA/CNRS UMR 8548, PSL Research University.

†CIIRC - Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague.

Method	Error [meters, degrees] [0.25, 5]	[0.50, 5]	[1.00, 10]	[2.00, 10]
DensePV [2]	38.9	55.6	69.9	74.2
DenseNV (cropped)	29.5	43.5	54.1	59.6
DenseNV	32.2	45.6	58.1	62.9
DensePNV (cropped)	39.5	56.8	71.7	**76.9**
DensePNV	**40.1**	**58.1**	**72.3**	76.6

Table A: **The impact of image cropping on pose verification, evaluated on the InLoc dataset [2].** We report the percentage of queries localized within given positional and rotational error bounds.

validate our design choice of preserving the field-of-view of the input images by avoiding cropping.

C. Semantic Superclass Construction

Below is the full mapping of the 150 semantic classes of the ADE20K dataset [3,4] to the five "superclasses" that we use to generate semantic masks. Each {} corresponds to one class from the CSAIL Semantic Segmentation pre-trained on MIT ADE20K dataset [3, 4] and each class is described by the labels inside the braces.

- *people:* {person, individual, someone, somebody, mortal, soul}
- *transient:* {plant, flora, plant, life}, {curtain, drape, drapery, mantle, pall}, {chair}, {mirror}, {rug, carpet, carpeting}, {armchair}, {seat}, {desk}, {lamp}, {cushion}, {base, pedestal, stand}, {box}, {grandstand, covered, stand}, {case, display, case, showcase, vitrine}, {pillow}, {screen, door, screen}, {flower}, {book}, {computer, computing, machine, computing, device, data, processor, electronic, computer, information, processing, system}, {swivel, chair}, {hovel, hut, hutch, shack, shanty}, {towel}, {apparel, wearing, apparel, dress, clothes}, {ottoman, pouf, pouffe, puff, hassock}, {bottle}, {plaything, toy}, {stool}, {barrel, cask}, {basket, handbasket}, {bag}, {cradle}, {ball}, {food, solid, food}, {trade, name, brand, name, brand, marque}, {pot, flowerpot}, {animal, animate, being, beast, brute, creature, fauna}, {bicycle, bike, wheel, cycle}, {screen, silver, screen, projection, screen}, {blanket, cover}, {sconce}, {vase}, {tray}, {ashcan, trash, can, garbage, can, wastebin, ash, bin, ash-bin, ashbin, dustbin, trash, barrel, trash, bin}, {fan}, {plate}, {monitor, monitoring, device}, {radiator}, {glass, drinking, glass}
- *stable:* {bed}, {cabinet}, {table}, {painting, picture}, {sofa, couch, lounge}, {shelf}, {wardrobe, closet, press}, {bathtub, bathing, tub, bath, tub}, {chest, of, drawers, chest, bureau, dresser}, {refrigerator, icebox}, {pool, table, billiard, table, snooker, table}, {bookcase}, {coffee, table, cocktail, table}, {bench}, {countertop}, {stove, kitchen, stove, range, kitchen, range, cooking, stove}, {arcade, machine}, {television, television, receiver, television, set, tv, tv, set, idiot, box, boob, tube, telly, goggle, box}, {poster, posting, placard, notice, bill, card}, {canopy}, {washer, automatic, washer, washing, machine}, {oven}, {microwave, microwave, oven}, {dishwasher, dish, washer, dishwashing, machine}, {sculpture}, {shower}, {clock}
- *fixed:* {wall}, {floor, flooring}, {ceiling}, {windowpane, window}, {door, double, door}, {railing, rail}, {column, pillar}, {sink}, {fireplace, hearth, open, fireplace}, {stairs, steps}, {stairway, staircase}, {toilet, can, commode, crapper, pot, potty, stool, throne}, {chandelier, pendant, pendent}, {bannister, banister, balustrade, balusters, handrail}, {escalator, moving, staircase, moving, stairway}, {buffet, counter, sideboard}, {stage}, {conveyer, belt, conveyor, belt, conveyer, conveyor, transporter}, {swimming, pool, swimming, bath, natatorium}, {step, stair}, {bulletin, board, notice, board}
- *outdoor:* {building, edifice}, {sky}, {tree}, {road, route}, {grass}, {sidewalk, pavement}, {earth, ground}, {mountain, mount}, {car, auto, automobile, machine, motorcar}, {water}, {house}, {sea}, {field}, {fence, fencing}, {rock, stone}, {signboard, sign}, {counter}, {sand}, {skyscraper}, {path}, {runway}, {river}, {bridge, span}, {blind, screen}, {hill}, {palm, palm, tree}, {kitchen, island}, {boat}, {bar}, {bus, autobus, coach, charabanc, double-decker, jitney, motorbus, motorcoach, omnibus, passenger, vehicle}, {light, light, source}, {truck, motortruck}, {tower}, {awning, sunshade, sunblind}, {streetlight, street, lamp}, {booth, cubicle, stall, kiosk}, {airplane, aeroplane, plane}, {dirt, track}, {pole}, {land, ground, soil}, {van}, {ship}, {fountain}, {waterfall, falls}, {tent, collapsible, shelter}, {minibike, motorbike}, {tank, storage, tank}, {lake}, {hood, exhaust, hood}, {traffic, light, traffic, signal, stoplight}, {pier, wharf, wharfage, dock}, {crt, screen}, {flag}

As detailed in Sec. 3.3 in the paper, we construct semantic masks by ignoring pixels belonging to the *people* and *transient* superclasses. This design choice was motivated by preliminary experiments with different ways to use semantic information. More precisely, we evaluated three variants of DensePV+S with semantic masks generated by the criteria listed below:

A We keep regions corresponding to the *stable* and *fixed* superclasses as informative and discard regions assigned to the other superclasses.

Method	Error [meters, degrees] [0.25, 5]	[0.50, 5]	[1.00, 10]	[2.00, 10]
DensePV [2]	38.9	55.6	69.9	74.2
DensePV+S (A)	**39.8**	57.4	**71.1**	**75.1**
DensePV+S (B)	39.2	56.2	70.5	**75.1**
DensePV+S (C)	**39.8**	**57.8**	**71.1**	**75.1**

Table B: **The impact of semantic masks, evaluated on the InLoc dataset [2].** We report the percentage of queries localized within given positional and rotational error bounds.

B We consider regions assigned to the superclass *people* as non-informative and regard all other regions as informative.

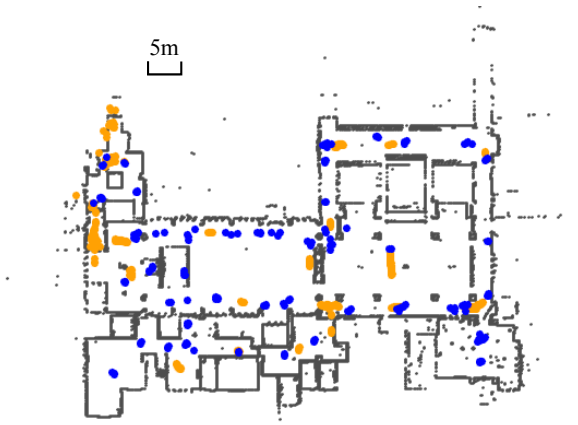
C We determine regions corresponding to the *people* and *transient* superclasses as non-informative and regard all other regions as informative.

Tab. B shows the comparison of DensePV [2] and DensePV+S with each type of semantic masking. All variations of DensePV+S considerably outperform the baseline. The best results are obtained with DensePV+S (C), which is the variant used in the paper.

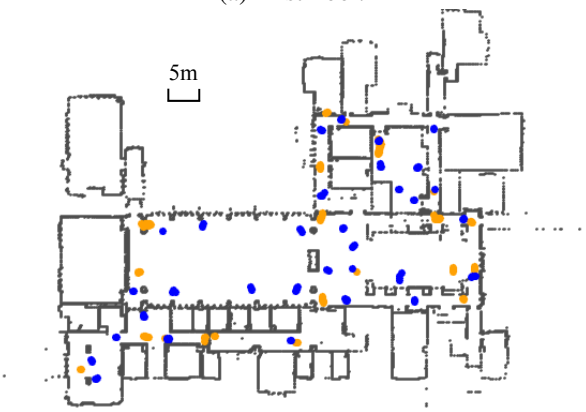
D. Training Data Generation

To train our learnable pose verification (*c.f*. Sec. 4), we use additional video sequences kindly provided by authors of [2], which were captured by them separately from the test images of the InLoc datasets. The images were captured using iPhone7 video streams in the same building as the InLoc dataset. In order to use the images for training, we created 6DoF ground-truth poses for them and used these poses to generate pose candidates for training. Fig. B shows the spatial distributions of the training images that we generated and manually verified. Note that there is little overlap between the original queries [2] and our training images, both in the first floor (a) and the second floor (b) of the building.

The ground-truth poses are computed as follows: **1)** From the original video sequences, we pick the frame with intervals of four seconds (key frame) and generate the manually verified 6DoF camera poses in a similar manner as the original InLoc dataset [2]. **2)** We additionally reconstruct the video frames adjacent to a key frame, using Structure-from-Motion (SfM) [1]. Note that in the bundle adjustment step, we fixed the 3D points that come from a feature in the key frame which has the depth information with respect to the database scans. This enables us to recover the scale of the SfM reconstruction. **4)** We visually inspect all poses and manually discard the images with incorrect poses. We also verify the reference poses by computing the overlap ratio with the relevant database scan with respect to the depth. We finally accepted 3,442 images that have more than 40% overlap. While training, 2,600 images in DUC1 (first floor) are used for training, and 842 images in DUC2 (second floor) are used for validation.



(a) First floor.



(b) Second floor.

Figure B: **Spatial distributions of the training images.** The orange dots in the figures show the camera positions of the training images, which we estimated and manually verified. The blue dots correspond to the positions of the original InLoc queries. Gray dots are the scanned 3D points of the InLoc dataset, showing the structure of the building.

E. Qualitative Results

Fig. C shows example localization results obtained by various methods on the InLoc dataset [2]. Fig. C (a) is an example on which the original DensePV [2] selects an inaccurate pose estimate, while our methods succeed when using the image-scan graph. Views rendered using only a single scan often cover only a part of the query view, which results in inaccurate pose verification, *i.e*., DensePV selects a query pose behind the wall. The image-scan graph enables us to use 3D points seen from the multiple scans related to the query. This results in a more complete synthesized image from a more accurate pose estimate, which is subsequently chosen by our approach.

Fig. C (b) is a typical scene on which DensePV with the scan-graph fails, while DensePV+S succeeds to accurately

localize the query. In this case, the query image is dominated by transient objects (shutter blinds) and people, which do not appear in the database images. Pose verification methods using only 3D structures (DensePV [2], DensePV w/ scan-graph) fail to achieve accurate localization in such scenes. DensePV+S discards the less-informative regions in the image based on semantic labels, which improves results. Using normals instead of semantic information has a similar effect in this scene.

The effectiveness of measuring surface normal consistency is shown in Fig. C (c). The query image shows a significant amount of weakly textured surfaces and regions of over-saturated pixels. Appearance-based pose verification methods (DensePV [2], DensePV w/ scan-graph) and our semantic-based DensePV+S approach fail to select an accurate pose candidate, since the scene appearance has largely changed between the query and the retrieved database image. On the other hand, DensePNV additionally compares surface normal directions, which provide useful information for this challenging query photo.

The benefit of combining semantics with surface normal consistency is shown in Fig. C (d). In the query, there are a number of transient objects, *e.g.*, chairs, movable tables, and people. This results in an inaccurate pose being selected by DensePNV since it directly computes surface normals even on such inconsistent objects. Using a semantic mask, DensePNV+S achieves better pose selection, ignoring those less informative regions.

References

[1] J. L. Schönberger and J.-M. Frahm. Structure-From-Motion Revisited. In *Proc. CVPR*, 2016. 3

[2] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii. InLoc: Indoor visual localization with dense matching and view synthesis. In *Proc. CVPR*, 2018. 1, 2, 3, 4, 5

[3] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ADE20K dataset. In *Proc. CVPR*, 2017. 2

[4] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ADE20K dataset. *IJCV*, 127(3):302–321, 2019. 2

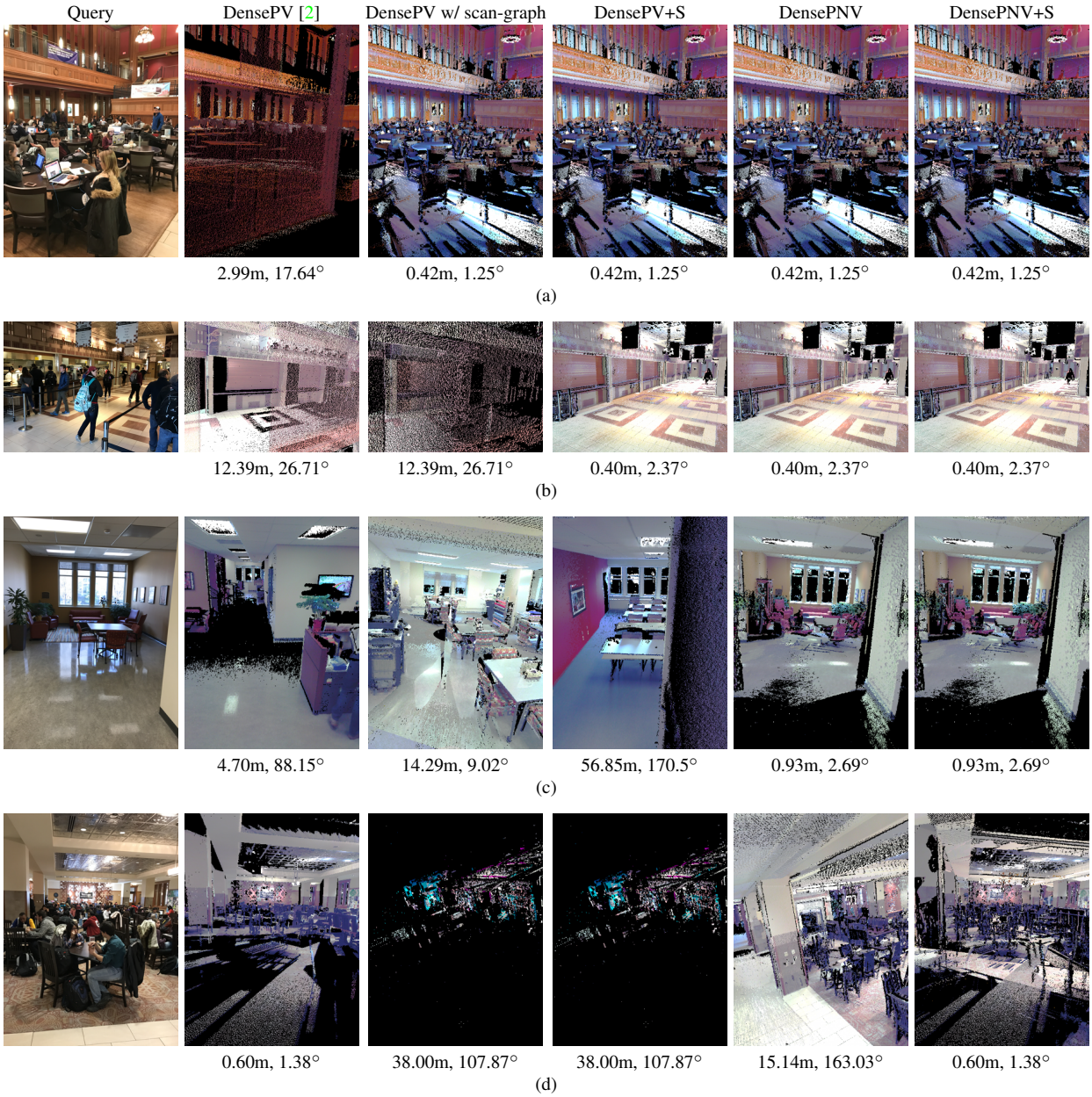


Figure C: **Qualitative examples of visual localization on the InLoc dataset [2].** Each row in the figure shows the query image (left) and the rendered views corresponding to the camera poses selected by different methods. The numbers under the synthesized images indicate the position and orientation errors with respect to the ground-truth poses. The scan-graph was used for the methods shown in columns 3 to 6.