VS-Net: Voting with Segmentation for Visual Localization - Supplementary Material -

Zhaoyang Huang[1,2*] Han Zhou[1*] Yijin Li[1] Bangbang Yang[1] Yan Xu[2]
Xiaowei Zhou[1] Hujun Bao[1] Guofeng Zhang[1] Hongsheng Li[2,3]

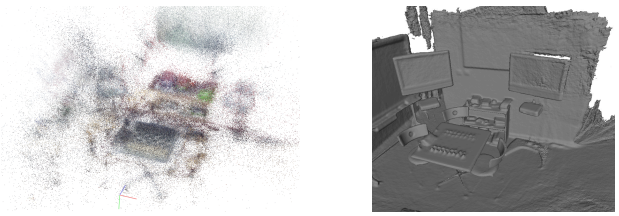
[1]State Key Lab of CAD&CG, Zhejiang University
[2]CUHK-SenseTime Joint Laboratory, The Chinese University of Hong Kong
[3]School of CST, Xidian University

Abstract

In this document, we provide details of our implementation complementing the original paper. Furthermore, we inspect the performance of our scene-specific landmarks and the voting-by-segmentation algorithm with extended experiments. Lastly, a video showing practical localization performance of our framework is given.

1. Detail of Implementation

Landmark generation. SuperVoxel [6] is a 3D over-segmentation algorithm that uniformly sets vast seeds in 3D space and raises a patch for each pre-defined seeds. We set a seed resolution, which can be approximately regarded as the size of patches, in each scene and discards the seeds that do not cover surfaces. Therefore, we can generate more landmarks with a smaller patch size. We present the patch size and the corresponding number of landmarks we use in the visual localization experiment in Tab. 1. In contrast with SfM-based visual localization frameworks relying on the SfM map, our scene-specific landmarks are computed from a 3D surface. We compare the SfM feature map built by COLMAP [9] and the 3D surface we use reconstructed by



(a) SfM Map (b) 3D Surface

Figure 1. Comparison of (a) SfM feature map and (b) TSDF surface.

Algorithm 1 Landmarks from Votes

INPUT:
$\mathbf{p}_i$: Pixel coordinates in current patch label
$\mathbf{d}_i$: Pixel votes at $\mathbf{p}_i$
$\hat{\mathbf{l}}_j^{(0)}$: Initial landmark of landmark j from RANSAC-based approach [7]
θ_{min}: Threshold of minimum neighbor pixels
θ_{dist}: Threshold of neighbor pixel distance
ϵ_{step}: Epsilon of landmark coordinate change
OUTPUT:
$\hat{\mathbf{l}}_j^{best}$: Best voting landmark
b_{valid}: Validity of the landmark

$\hat{\mathbf{l}}_j^{best} \leftarrow \hat{\mathbf{l}}_j^{(0)}$
$b_{valid} \leftarrow$ **true**
$t \leftarrow 1$
while $t <$ Max Iteration **do**
  $S = \{\|\mathbf{p}_i - \hat{\mathbf{l}}_j^{(t-1)}\|_2 < \theta_{dist}\}$
  if $|S| < \theta_{min}$ **then**
    $b_{valid} \leftarrow$ **false**
    break
  end if
  for all $\mathbf{p}_i \in S$ **do**
    Compute normal of voting map $\mathbf{n}_i \leftarrow \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \mathbf{d}_i$
  end for
  $\hat{\mathbf{l}}_j^{(t)} \leftarrow \left(\sum_S \mathbf{n}_i \mathbf{n}_i^\top\right)^{-1} \left(\sum_S \mathbf{n}_i \mathbf{n}_i^\top \mathbf{p}_i\right)$
  if $||\hat{\mathbf{l}}_j^{(t)} - \hat{\mathbf{l}}_j^{(t-1)}||_2 < \epsilon_{step}$ **then**
    break
  end if
  $\hat{\mathbf{l}}_j^{best} \leftarrow \hat{\mathbf{l}}_j^{(t)}$
  $t \leftarrow t + 1$
end while

	7Scenes							Cambridge Landmarks				
	Ches.	Fire	Head.	Offi.	Pump.	Kitc.	Stair.	College	Court	Hospital	Shop	Church
Size	15cm	10cm	5cm	15cm	15cm	15cm	10cm	2.0m	3.0m	1.5m	0.8m	1.2m
Num.	4330	5052	10058	4982	4702	6402	13878	4081	2799	2940	4657	4141

Table 1. The number of landmarks in the visual localization experiments.

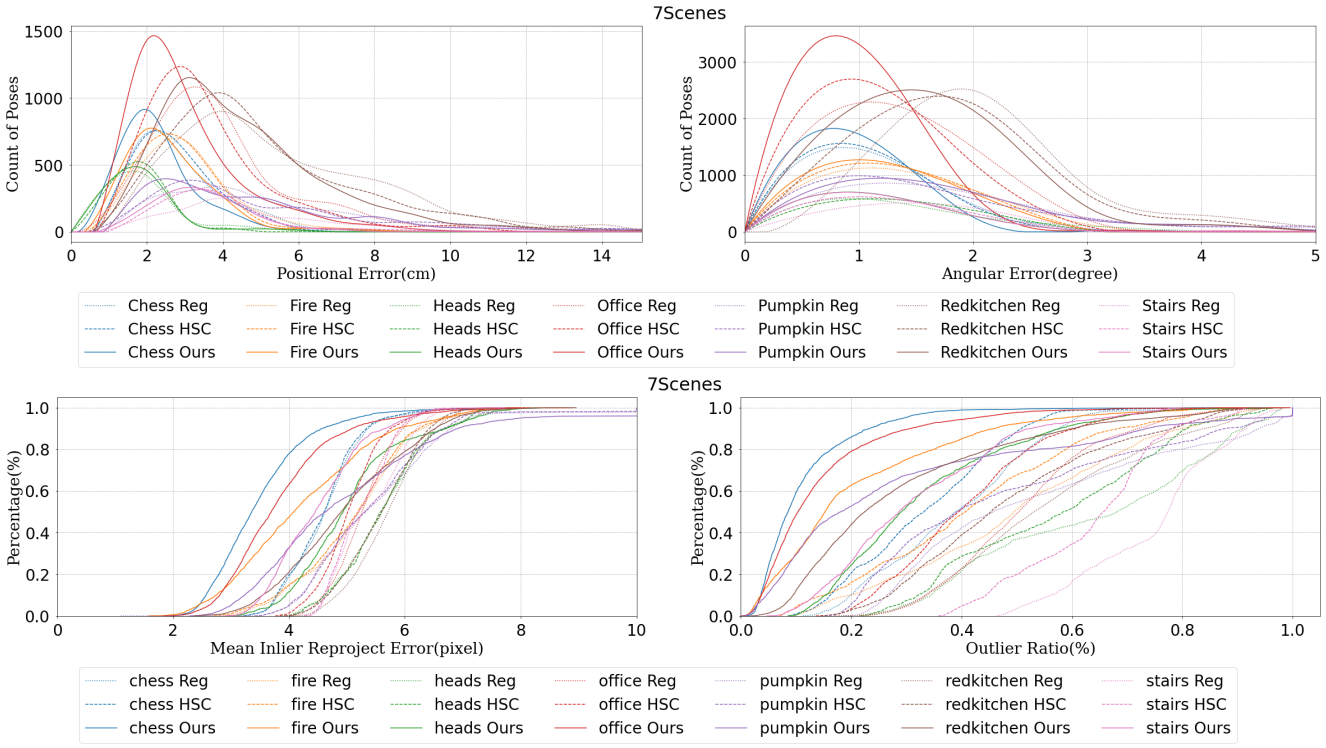


Figure 2. Scene-specific landmarks vs. Reg vs. HSC-Net in individual scenes.

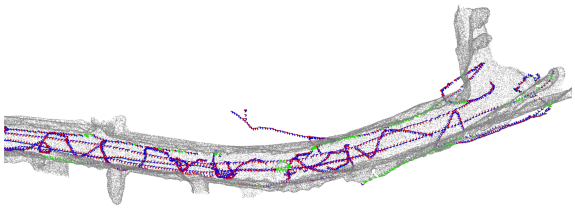


Figure 3. The reconstructed map (gray points) and camera trajectory (purple) of *Street* in the *Cambridge Landmarks Dataset*.

KinectFusion [5] in Fig. 1. The SfM feature map is typically messy and not accurate enough even in a static indoor scene, while the 3D surface is more tidy and accurate, which benefits the following localization.

VS-Net details. We use the DeepLabv3 [1] as the backbone for landmark segmentation and add a decoder for pixel-wise voting. Our models are trained by Adam optimizer with a learning rate of 1e-4, a weight decay of 5e-4, and a batch size of 2. The images are respectively scaled to 640×480 and 960×540 in the 7-Scenes dataset and the Cambridge Landmarks dataset. The length of prototype embedding used in the 7-Scenes dataset and the Cambridge Landmarks dataset are 24 and 64. k is set as 2 in the kNN search of prototype-based triplet loss. We use affine transformation data augmentation in training with the same parameter settings in hsc-net [3] on both of these datasets. We elaborate the EM-like landmark location refinement algorithm in Alg. 1.

Case study of *Street*. There are six scenes in the *Cambridge Landmarks Dataset*. Following previous works [3], we only compare VS-Net with others in five scenes except *Street* because the provided poses in *Street* is not reasonable. We present a part of the 3D map and the camera trajectory computed from official ground-truth camera poses of the *Street* in Fig. 3. The cameras within 0.5m to the wall are rendered as green. As we can see, the camera trajectory drifts seriously.

2. Extended Experiments

Scene-specific landmarks vs. Reg vs. HSC-Net in individual scenes. We compare scene-specific landmarks and scene coordinates in individual scenes. Reg [3] is the base-

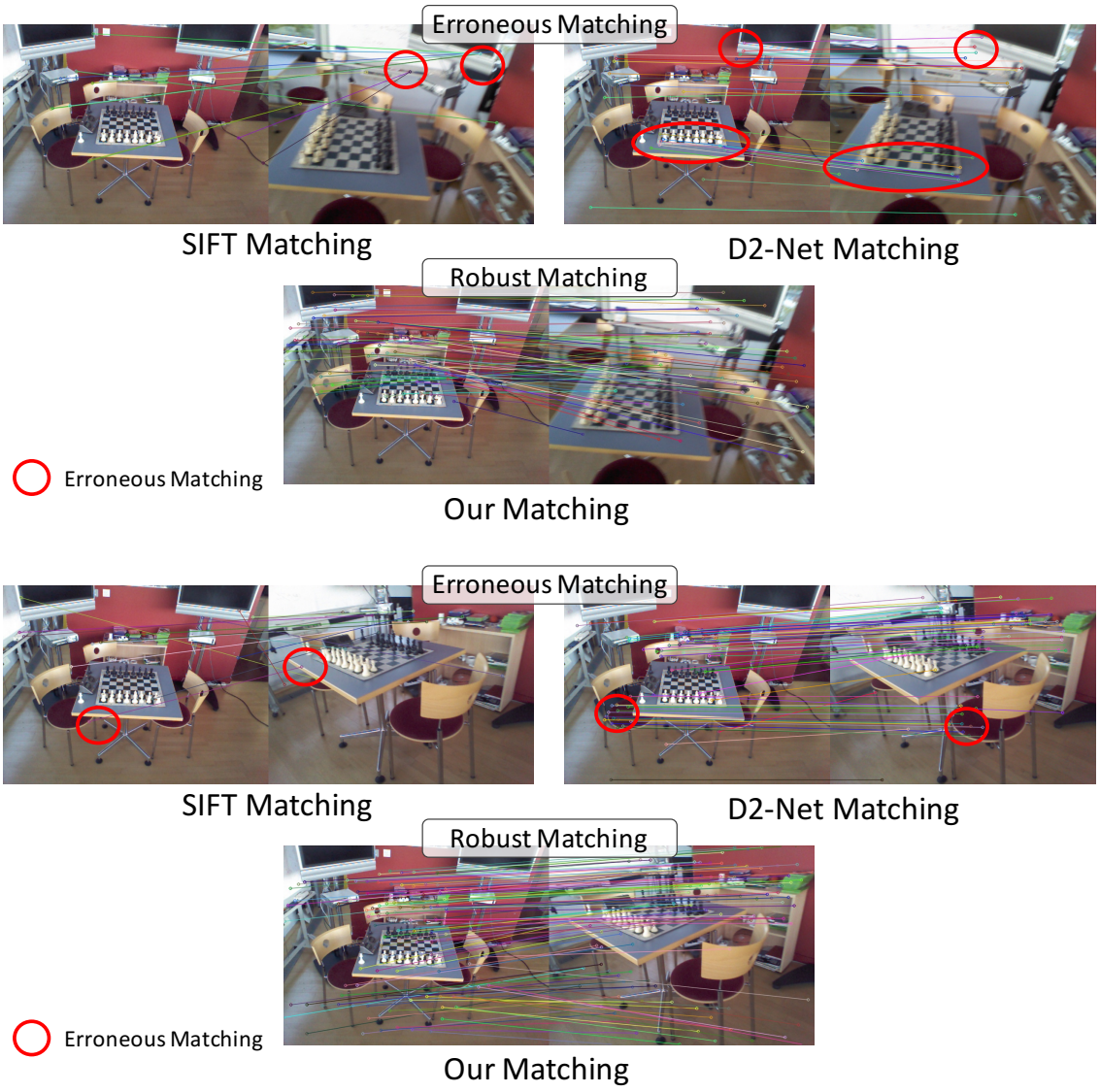


Figure 4. Feature matching with fundamental matrix based RANSAC.

line in scene coordinate regression and HSC-Net [3] is the state-of-the-art scene coordinate regression method. We compare VS-Net with them to show the superiority of our proposed scene-specific landmarks. In addition to the pose accuracy metrics, we measure the robustness and accuracy of correspondences by two other metrics: the outlier ratio and mean reprojection error of inliers. We compute the reprojection error of scene-specific landmarks and scene coordinates with ground-truth poses. The established 2D-to-3D correspondences whose reprojection errors are lower than 10 pixels are regarded as inliers, and the others are treated as outleirs. In Fig. 2, the mean inlier reprojection error and outlier ratio of scene-specific landmarks are concentrated on lower error intervals, which means that the accuracy and robustness of scene-specific landmarks are better than scene coordinates in most query images. The number of poses computed from scene-specific landmarks consistently surpasses scene coordinates at the low threshold in both positional error and angular error, which indicates that scene-specific landmarks exhibit better pose estimation performance in most scenes.

Fundamental-matrix based matching. In the view of feature-based visual localization, landmark labels can be regarded as the feature descriptors that associate two detected landmarks. To show that VS-Net generates more robust correspondences in a specified scene, we compare our landmark association with SIFT [4], D2-Net [2] feature matching in Figs. 4. All the matches have been filtered by epipolar

geometry with RANSAC. R2D2 [8] is not able to figure out more than three correspondences, which can not be applied to the following fundamental matrix estimation, so we do not present it in the figures. The images are extracted from the Seq-3 in the 7-Scenes chess scene, which is a test sequence. The image pairs are constituted by the first frame and a frame sampled from the subsequent frames. Even after RANSAC, there are many erroneous matches in SIFT and D2-Net when the baseline is large because the textures are locally similar. We highlight some of them with red circles. By contrast, our landmark association keeps good robustness and accuracy under such a significant viewpoint changing.

Visual localization in practical environments We provide a supplementary video that qualitatively demonstrates the performance of VS-Net in severe practical environments. The Cambridge Landmarks dataset provides some challenging images and videos while they are not contained in the test set. We compute the camera poses of the images through the proposed visual localization framework and project the model into the views of the original sequences for performance evaluation. The successfully detected and associated landmarks are visualized with red crosses for better understanding. The results are presented in the supplementary video, which indicates the prominent robustness of our framework.

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. 2

[2] Mihai Dusmanu, Ignacio Rocco, Tomás Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable CNN for joint description and detection of local features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8092–8101, 2019. 3

[3] Xiaotian Li, Shuzhe Wang, Yi Zhao, Jakob Verbeek, and Juho Kannala. Hierarchical scene coordinate classification and regression for visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11983–11992, 2020. 2, 3

[4] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 3

[5] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew W Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, volume 11, pages 127–136, 2011. 2

[6] Jeremie Papon, Alexey Abramov, Markus Schoeler, and Florentin Wörgötter. Voxel cloud connectivity segmentation - supervoxels for point clouds. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, Portland, Oregon, June 22-27 2013. 1

[7] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019. 1

[8] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. In *Advances in Neural Information Processing Systems*, pages 12405–12415, 2019. 4

[9] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1