SCENE-UNIFIED IMAGE TRANSLATION FOR VISUAL LOCALIZATION

Sheng Han, Wei Gao, Yiming Wan and Yihong Wu

{sheng.han, wgao, yiming.wan, yhwu}@nlpr.ia.ac.cn
NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100049, China

ABSTRACT

Visual localization is a key technology in the field of 3D robot vision. One of its major difficulties lies in how to deal with the challenges brought by the appearance changes of query images and database images caused by large time spans. Many methods focus on extracting more robust features from images to deal with the impact of complex scenes. In this paper, we explore the impact of image translation on visual localization tasks in complex scenes. We propose UniGAN - a modified image translation model, fusing semantic label constraints and finer reconstruction losses, to unify images captured under different environmental conditions to a standard scene more suitable for localization tasks. To estimate the 6-DOF camera pose, a two-stage localization framework composed of image retrieval and local matching is utilized. Experiments show that our method outperforms the state-of-the-art in terms of both accuracy and robustness to environmentally sensitive scenes.

***Index Terms*—** Visual Localization, Image Translation, Semantic Segmentation, GANs

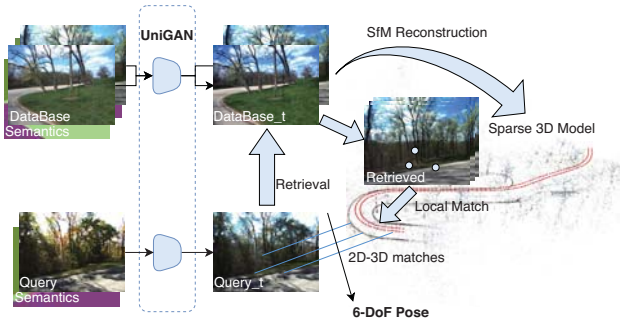

Fig. 1: **Scene-unified Localization**. The query and database images with semantic information are first translated into a unified scene by UniGAN, then the translated database images are used to build a 3D point cloud model. With the sparse 3D model, a two-stage process involving image retrieval and local matching can be carried out to obtain the 6-DoF pose of the query image. This pipeline is demonstrated efficient to cope with the challenges of visual localization in long-span scenes.

1. INTRODUCTION

Visual localization is a key technology in 3D computer vision for its important applications in visual tasks such as Augmented Reality[1], 3D reconstruction[2], SLAM, and assistance system for self-driving cars. The core objective of visual localization is to estimate the 6-Degree-of-Freedom(6-DoF) pose of a camera with respect to a global coordinate system. To this end, structure-based approaches like Active Search [3] and City-Scale Localization (CSL) [4] focus on establishing 2D-3D matches between features in a query image and the 3D scene model, while image-based methods such as DenseVLAD [5], NetVLAD [6], and FAB-MAP [7] utilize the extracted global descriptors for retrieval, and then the pose of the nearest image retrieved is regarded as the pose of the query image. Obviously, both solutions rely heavily on feature extraction of the captured images, which vary greatly over a large time span. As the application area continues to expand, advanced algorithms with high accuracy and robustness are called, especially in challenging scenes with large variations in illumination, weather, or seasons.

At this stage, the development of deep learning provides new ideas for many computer vision problems. Many tasks, such as image classification, object detection, semantic segmentation, image translation, etc., are greatly facilitated by neural networks. Thanks to these outstanding works that made it possible to use deep learning for visual localization, many studies start integrating deep learning into localization systems to tackle challenging scenes. A common application is to extract feature points that are more robust to environmental factors than hand-crafted ones through a convolutional neural network (CNN) [8, 9]. Another typical representative is PoseNet [10], which regresses camera pose directly with an end-to-end approach. This process can obtain relatively high operating efficiency as it avoids the registration from 2D images to 3D point clouds. Nevertheless, since the data calibration of PoseNet is based on SfM, its precision is also

This work was supported in part by the National key R&D Program of China (2016YFB0502002) and in part by the Natural Sciences Foundation of China (61872361, 61836015). Corresponding author: Wei Gao (wgao@nlpr.ia.ac.cn)

limited by the scene similarity of query image and image set used to build the 3D model. In addition, end-to-end methods have not yet reached the same level of accuracy as traditional methods.

Unlike other methods to improve the robustness of the algorithm, inspired by image translation, we try to solve the problem of dissimilar scenes by translating the query image and database images into a unified scene. As a pre-processing for a two-stage process involving image retrieval and local matching, this technique outperforms all current state-of-the-art methods in exact 6-DoF localization. In general, our contributions can be summarized as follows:

- We set a new state-of-the-art in the public benchmark for large-scale localization with better robustness in challenging scenes.
- We demonstrate the practical usefulness of scene unification in tackling luminosity and season changes in visual localization.
- We introduce UniGAN, an image translation generation adversarial network incorporating semantic labels for unifying challenging scenes.

2. RELATED WORK

Image translation. Many visual tasks require converting images from one domain into another, and that leads Isola et al. [11] to propose the first unified framework for image translation based on conditional generative adversarial networks (Conditional GANs). Since aligned image pairs as training data are difficult to obtain, image translation trained on unpaired images has been addressed for various domain translations such as cats to dogs or summer to winter [12]. Anoosheh et al. [13] recently extend the translation between two domains to multiple domains like numerous artistic styles or four seasons. In addition, the UNIT framework proposed by Liu et al. [14] gives the assumption that corresponding images in two domains share the same latent space, which is adopted in our UniGAN to get the content code of images from different scenes.

Visual localization. The 6-DoF visual localization methods have long been classified as either structure-based or image-based. Although there have been some learning-based approaches [10, 15] directly regress the camera pose from a single image, they are not competitive in term of accuracy. Structure-based methods [4, 3] use database images to build a 3D point cloud, and directly match the key points of a query image with the 3D points in the SfM structure. Then the camera pose can be estimated from the resulting 2D-3D matches by solving a Perspective-n-Point (PnP) problem [16] within the RANSAC scheme [17]. Nevertheless, the 2D-3D matches may be computationally complex or ambiguous in scenes with large scale or strong appearance changes. Image-based methods [5, 6, 7] are reasonably more effective and robust than structure-based ones, since they approximate the pose of a query image with the most similar photo retrieved from the image database, but that's also why they are not sufficiently precise.

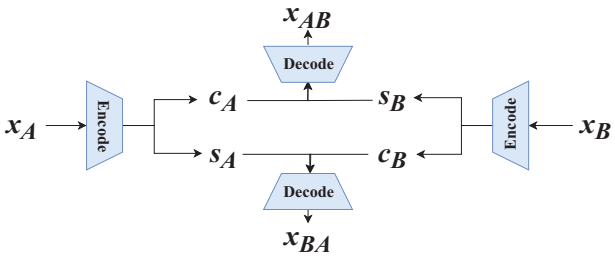

Fig. 2: **Cross-domain Image Translation**.

To combine the efficiency and precision of approaches mentioned above, hierarchical localization methods [18] try to split the process into a global coarse image retrieval followed by a local fine pose estimation. In this paper, we follow this strategy.

3. PROPOSED METHOD

As shown in Figure 1, our method to estimate the accurate camera pose of a query image with a great difference from database images in scene conditions like weather, seasons and illumination, etc. is as follows. First, we define the scene when capturing the database as domain B, and various other scenes with different environmental conditions, where the query images are, as domain A. An image translation model named UniGAN is trained to translate between domain A and domain B. Second, the generator $G_{A\to B}$ is used to translate images from both domain A and B to a unified representation in the standard domain U. The translated database images are then utilized to build the unified sparse 3D model through SfM. Then the accurate 6-DoF pose of the query image is estimated with a two-stage process involving global retrieval and local match.

3.1. Semantic Segmentation

Semantic segmentation is a preprocessing for training our UniGAN. We employ the architecture of Cross-season-segmentation [19], which utilizes 2D-2D point matches between images taken during different image conditions to train a convolution neural network for semantic segmentation. The output of Cross-season-segmentation are RGB semantic graphs with corresponding colors mapped, for avoiding the influence of color selection on the training process of UniGAN, we use RGBS images, which consist of original RGB images(RGB) and semantic label graphs(S) produced before

the color mapping in Cross-season-segmentation, as the input of UniGAN.

3.2. UniGAN

Our image-translation model is built using MUNIT [12] as the base. We accept the setting that an image x can be encoded into a content code c and a style code s. And image translation is essentially an autoencoder as shown in Figure 2. We set an encoder E_i and a decoder G_i for each domain X_i ($i = A, B$), where $E_i(x_i) = (E_i^c(x_i), E_i^s(x_i)) = (c_i, s_i)$. Image translation is performed by swapping encoder-decoder pairs.

Our loss function includes a bidirectional reconstruction loss, a cycle consistency loss, and an adversarial loss. For bidirectional reconstruction loss, the model should be able to reconstruct the image x_i in image $\rightarrow$ lantent $\rightarrow$ image direction, and its lantent codes (c_i, s_i) should be able to be reconstructed in lantent $\rightarrow$ image $\rightarrow$ lantent direction. The loss terms are defined as follows:

$$\begin{aligned} L_{recon}^{x_A} =& \mathbb{E}[(1-\alpha)\cdot\|G_A(E_A^c(x_A), E_A^s(x_A)) - x_A\|_1 \\ &+ \alpha\cdot\text{MS-SSIM}(G_A(E_A^c(x_A), E_A^s(x_A)), x_A)] \end{aligned} \tag{1}$$

$$L_{recon}^{c_A} = \mathbb{E}[\|E_B^c(G_B(c_A, s_B)) - c_A\|_1] \tag{2}$$

$$L_{recon}^{s_A} = \mathbb{E}[\|E_A^s(G_A(c_B, s_A)) - s_A\|_1] \tag{3}$$

The other loss terms $L_{recon}^{x_B}$, $L_{recon}^{c_B}$ and $L_{recon}^{s_B}$ are defined in a similar manner. We use a combination of L_1 loss and MS-SSIM loss for image reconstruction according to Zhao's work [20] to preserve colors, luminance and the contrast in high-frequency regions. For convenience, we next use the function

$$\begin{aligned} &R^{Mix}(m, n) = \\ &\quad\mathbb{E}[(1-\alpha)\cdot\|m-n\|_1 + \alpha\cdot\text{MS-SSIM}(m, n)] \end{aligned} \tag{4}$$

to concisely represent the reconstruction loss between image m and n.

To ensure that translated images generated by our model are indistinguishable from real images in the target domain, the cycle consistency loss and adversarial loss are introduced as follows:

$$\begin{aligned} L_{cyc} =& R^{Mix}(G_{B\rightarrow A}(G_{A\rightarrow B}(x_A)), x_A) \\ &+ R^{Mix}(G_{A\rightarrow B}(G_{B\rightarrow A}(x_B)), x_B) \end{aligned} \tag{5}$$

$$\begin{aligned} L_{GAN}^{x_B} =& \mathbb{E}[\log(1 - D_B(G_{A\rightarrow B}(x_A)))] \\ &+ \mathbb{E}[\log D_B(x_B)] \end{aligned} \tag{6}$$

where D_B is the discriminator that tries to distinguish between translated images and real images in domain B, and $G_{A\rightarrow B}$ is the generator that translates images from domain A into domain B, which is equivalent to $G_B(c_A, s_B)$. The discriminator D_A, the generator $G_{B\rightarrow A}$ and loss $L_{GAN}^{x_A}$ are defined similarly.

We use a weighted sum of the bidirectional reconstruction loss, cycle consistency loss and adversarial loss mentioned above as the total loss of our UniGAN. Hence the training objective is

$$\begin{aligned} &\min_{E_A,E_B,G_A,G_B}\max_{D_A,D_B} L(E_A, E_B, G_A, G_B, D_A, D_B) = \\ &L_{GAN}^{x_A} + L_{GAN}^{x_B} + \lambda_{cyc}L_{cyc} + \lambda_x(L_{recon}^{x_A} + L_{recon}^{x_B}) + \\ &\lambda_c(L_{recon}^{c_A} + L_{recon}^{c_B}) + \lambda_s(L_{recon}^{s_A} + L_{recon}^{s_B}) \end{aligned} \tag{7}$$

where λ_{cyc}, λ_x, λ_s, λ_c are weights that control the importance of reconstruction terms.

3.3. Localization Process

We follow the hierarchical localization methods mentioned in Section 2 to estimate the 6-DoF pose. As preprocessing, both query images and database images are translated with the generator $G_{A\rightarrow B}$ to domain U. Then the global descriptors of translated images are extracted by NetVLAD [6], and the image retrieval follows to obtain 10 prior images from the translated database for a query image. After that, the local descriptors, Superpoints [9], are used to get the 2D-3D matches between the query image and SfM models related to prior images. Finally, the accurate 6-DoF pose of a query image is obtained by solving a PnP problem within the RANSAC scheme.

4. EXPERIMENTS

4.1. Datasets and Training Setup

We evaluate our work on the CMU Seasons Dataset [21]. It contains 17 slices that cover three types of scenery (urban, suburban and park). It contains 7,159 database images and 75,335 query images record in different weather and seasons. The images were collected using a rid of two cameras mounted at 45 degrees forward/left and forward/right angles on the roof of an SUV, traversing an 8.5 km long route. We use the park scenery (slice11-17) to train our UniGAN and the whole dataset to evaluate it. In our experiment, the images are resized to 360×360 for training and kept the original size (1024×768) when testing. We train the model for a maximum of 500,000 iterations with a batch size of 1 on one NVIDIA GTX1080 Ti using PyTorch.

4.2. Results Analysis

To compare how the use of translated images at different stages of the localization process can affect the accuracy, we test the results of different data configurations in the two-stage localization framework NV+SP (NetVLAD+Superpoint). Due to space limitations, we cannot detail the comparison accuracy, but the conclusion is, the accuracy reaches the highest when using global descriptors extracted from translated query images and translated database images during image retrieval, while

Table 1: **Evaluation of the location** on the CMU Seasons dataset. We report the recall [%] at different distance and orientation thresholds and highlight for each of them the **best** and **second-best** methods. X+Y denotes hierarchical localization with X (Y) as global (local) descriptors.

	CMU					
	foliage	mixed foliage	no foliage	urban	suburban	park
distance [m]	.25/.50/5.0	.25/.50/5.0	.25/.50/5.0	.25/.50/5.0	.25/.50/5.0	.25/.50/5.0
orient. [deg]	2/5/10	2/5/10	2/5/10	2/5/10	2/5/10	2/5/10
Active Search [3]	28.8 / 32.5 / 35.9	25.1 / 29.4 / 33.9	52.5 / 59.4 / 66.7	55.2 / 60.3 / 65.1	20.7 / 25.9 / 29.9	12.7 / 16.3 / 20.8
CSL [4]	16.3 / 19.1 / 26.0	15.2 / 18.8 / 28.6	36.5 / 43.2 / 57.5	36.7 / 42.0 / 53.1	8.6 / 11.7 / 21.1	7.0 / 9.6 / 17.0
DenseVLAD [5]	13.2 / 31.6 / 82.3	16.2 / 38.1 / 85.4	17.8 / 42.1 / 91.3	22.2 / 48.7 / 92.8	9.9 / 26.6 / 85.2	10.3 / 27.0 / 77.0
NetVLAD [6]	10.4 / 26.1 / 80.1	11.0 / 26.7 / 78.4	11.8 / 29.1 / 82.0	17.4 / 40.3 / 93.2	7.7 / 21.0 / 80.5	5.6 / 15.7 / 65.8
FABMAP [7]	1.1 / 2.7 / 16.5	1.0 / 2.5 / 14.7	3.6 / 7.9 / 30.7	2.7 / 6.4 / 27.3	0.5 / 1.5 / 13.6	0.8 / 1.7 / 11.5
LocalSfM [21]	55.4 / 57.0 / 59.9	52.4 / 55.1 / 58.6	70.8 / 72.7 / 75.9	72.8 / 74.1 / 76.1	55.2 / 57.7 / 61.3	41.8 / 44.5 / 48.7
NV+SP [18]	69.4 / 74.9 / 87.2	73.5 / 80.1 / 89.3	**83.7** / **88.5** / **93.1**	91.7 / 94.6 / **97.7**	**74.6** / **81.6** / **91.4**	54.3 / 62.5 / 79.0
UniGAN(RGB)+NV+SP(Ours)	**71.3** / **76.6** / **88.2**	**74.5** / **80.8** / **90.3**	81.9 / 87.0 / 92.9	**92.1** / **94.8** / **98.0**	**75.9** / **81.6** / 90.8	**55.5** / **64.1** / **81.0**
UniGAN(RGBS)+NV+SP(Ours)	**71.7** / **76.9** / **88.4**	**75.2** / **81.5** / **90.8**	**82.9** / **87.9** / **93.4**	**92.4** / **95.0** / **98.0**	**75.9** / **82.1** / **91.0**	**56.8** / **65.1** / **81.7**

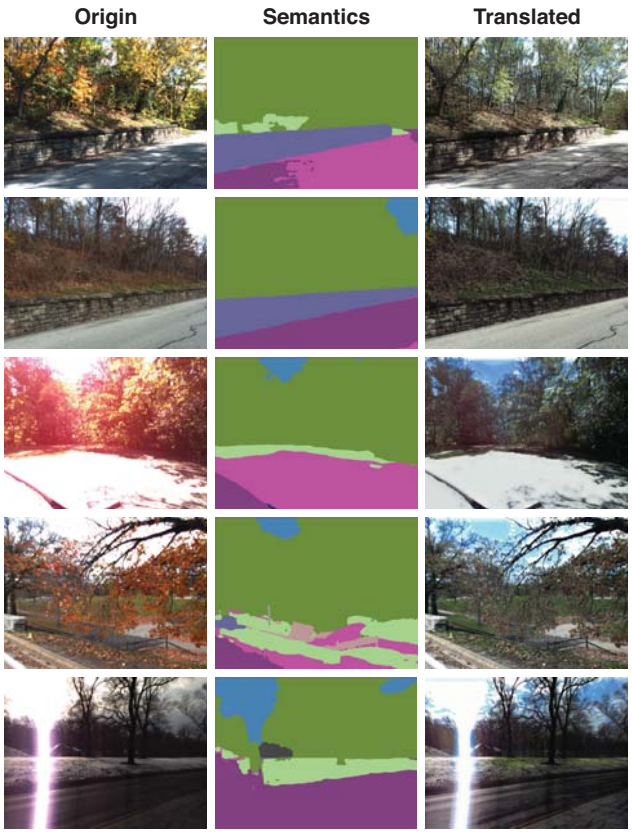

Fig. 3: **Visual Comparison**. Left to right column: Example origin images before translation, semantic graphs obtained by segmentation, translated images obtained by UniGAN.

using superpoints extracted from translated query images and translated database images during local matching.

On this basis, we repeated the evaluation with our UniGAN under the optimal configuration. In the first experiment, we use the original RGB images for training and testing. The result in Table 1 shows that the unified scene is beneficial to improve the localization accuracy. In the second experiment, we replace the RGB images with RGBS images mentioned in Section 3.1, and the result in Table 1 shows further improvement in accuracy with the introducing of semantics. Figure 3 shows the visual comparison between images before and after the translation, and Table 1 indicates our model indeed exceeds all other methods in the public benchmark for large-scale localization [21], especially in the high-precision threshold $(0.25m, 2°)$ and scenes that are greatly affected by seasonal factors.

5. CONCLUSION

In this paper, we propose a visual localization system based on the image translation model UniGAN. We introduce semantic label constraints and finer reconstruction losses during model training, and for the first time use the translated images for 3D scene model construction. Our results show that our approach significantly outperforms state-of-the-art works on the challenging task of localizing multi-scenario queries against a set of database images captured in a specific scene.

6. REFERENCES

[1] Robert Castle, Georg Klein, and David W Murray, "Video-rate localization in multiple maps for wearable augmented reality," in *2008 12th IEEE International Symposium on Wearable Computers*. IEEE, 2008, pp. 15–22.

[2] Johannes L Schonberger and Jan-Michael Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4104–4113.

[3] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla, "Are large-scale 3d models really necessary for accurate visual localization?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1637–1646.

[4] Linus Svärm, Olof Enqvist, Fredrik Kahl, and Magnus Oskarsson, "City-scale localization for cameras with known vertical direction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 7, pp. 1455–1461, 2016.

[5] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1808–1817.

[6] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[7] Mark Cummins and Paul Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[8] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler, "D2-net: A trainable cnn for joint detection and description of local features," *arXiv preprint arXiv:1905.03561*, 2019.

[9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 224–236.

[10] Alex Kendall, Matthew Grimes, and Roberto Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[12] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172–189.

[13] Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool, "Combogan: Unrestrained scalability for image domain translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 783–790.

[14] Ming-Yu Liu, Thomas Breuel, and Jan Kautz, "Unsupervised image-to-image translation networks," in *Advances in neural information processing systems*, 2017, pp. 700–708.

[15] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz, "Geometry-aware learning of maps for camera localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2616–2625.

[16] Laurent Kneip, Davide Scaramuzza, and Roland Siegwart, "A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation," in *CVPR 2011*. IEEE, 2011, pp. 2969–2976.

[17] Martin A Fischler and Robert C Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[18] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12716–12725.

[19] Mans Larsson, Erik Stenborg, Lars Hammarstrand, Marc Pollefeys, Torsten Sattler, and Fredrik Kahl, "A cross-season correspondence dataset for robust semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9532–9542.

[20] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 47–57, 2016.

[21] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al., "Benchmarking 6dof outdoor visual localization in changing conditions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8601–8610.