Danish Airs and Grounds: A Dataset for Aerial-to-Street-Level Place Recognition and Localization

Andrea Vallone[1*], Frederik Warburg[1*], Hans Hansen[2], Søren Hauberg[1] and Javier Civera[3]

***Abstract*—Place recognition and visual localization are particularly challenging in wide baseline configurations. In this paper, we contribute with the *Danish Airs and Grounds* (DAG) dataset, a large collection of street-level and aerial images targeting such cases. Its main challenge lies in the extreme viewing-angle difference between query and reference images with consequent changes in illumination and perspective. The dataset is larger and more diverse than current publicly available data, including more than 50 km of roads in urban, suburban and rural areas. All images are associated with accurate 6-DoF metadata that allows the benchmarking of visual localization methods. Additionally, we validate our data by presenting the results of a simple map-to-image re-localization baseline. that first estimates a dense 3D reconstruction from the aerial images and then matches query street-level images to street-level renderings of the 3D model. The dataset can be downloaded at: https://frederikwarburg.github.io/DAG/.**

***Index Terms*—Localization, Mapping, Data Sets for SLAM, Deep Learning for Visual Perception, Visual Learning**

I. Introduction

Estimating the 6-Degrees-of-Freedom (6-DoF) camera pose in a known scene map is a core component in many applications such as autonomous driving, robotics, and augmented reality. Visual localization pipelines are typically divided into two stages. First, place recognition methods obtain a coarse camera pose by finding images from a large database of registered images that are *similar* to a given query image. Second, a visual localization method refines the relative pose between the retrieved and the query images, in most cases relying on feature extraction and matching.

Handcrafted descriptors have shown impressive performance for both place recognition and visual localization (*e.g.*, [1], [2]), but are limited to small changes in perspective, illumination and scene structure. In the last decade, learning-based feature extractors and descriptors have overcome these limitations, even for drastic appearance changes such as day-to-night or summer-to-winter. The need for training data and fair benchmarking have motivated the release of many large and challenging place recognition [3], [4], [5] and localization [6], [7], [8] datasets that focus especially on appearance and viewpoint changes. Following this trend, aerial mapping is a particularly interesting application to study viewpoint invariances.

Fig. 1: Sample images from our DAG dataset illustrating the challenge of retrieving the closest aerial image given a street-level query, and of registering the query in the aerial reference frame.

Moreover, aerial mapping has a wide range of applications. Compared to street-view mapping, in which drivers or pedestrians have to traverse every road, aerial images provide a more scalable method for mapping large areas. The alignment of multiple mapping sequences is simpler with airplane photos than street-level sequences, because of the large receptive field and overlap of aerial images. Compared to satellite photos, airplanes provide oblique views and higher resolution that allows for detailed mapping of building facades (*e.g.* see the detailed texture on the facade of the power plant in Fig. 1).

This paper contributes to the ongoing research on visual place recognition and localization with a challenging dataset presenting extreme viewpoint changes. Specifically, the *Danish Airs and Grounds* (DAG) dataset targets visual place recognition and localization between aerial and street-level images. DAG contains diverse urban and suburban environments and is currently the largest and most diverse dataset of its kind.

To validate the dataset, we implemented a baseline method for aerial-to-street-level visual localization. We first create a 3D model from aerial images from which we render street-level images, thus reducing the view-angle difference between query and database images. Similarly to [9], we find

Manuscript received: Feb, 3, 2022; Revised May, 11, 2022; Accepted June, 9, 2022.

This paper was recommended for publication by Editor Cesar Cadena Lerma upon evaluation of the Associate Editor and Reviewers' comments.

* The first two authors contributed equally to the work. [1]Technical University of Denmark, {s192327,frwa,sohau}@dtu.dk, [2]Dansk Drone Kompagni ApS, hans@dronekompagniet.dk, [3]I3A, Universidad de Zaragoza, jcivera@unizar.es

This work was funded by the Spanish Government (grant PGC2018-096367-B-I00), the Aragón Government (grant DGA FSE-T45 20R), research grants (15334, 42062) from VILLUM FONDEN, and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 757360).

Digital Object Identifier (DOI): see top of this page.

that pre-trained feature descriptors are effective for visual localization between the rendered and the query images. Our method, however, comes at the expense of an expensive 3D reconstruction and rendering processes. We hope that the release of the DAG dataset will facilitate research in direct visual localization between aerial and street-level images without the need of rendered views, and that our implementation will contribute as a valuable baseline.

II. Related Work

Visual localization pipelines are typically divided into place recognition and 6-DoF localization. In this section, we review the main works in both stages, the most relevant datasets for visual localization and specific works in aerial-to-street-level localization.

A. Visual Place Recognition

Visual place recognition is often cast as an image retrieval task, where the goal is to find images from the same place as a query image in a large database of geo-registered images. The definition of same-place varies, but usually two places are considered the same if they are within a certain distance radius (25 meters is a common choice). Retrieval methods are more scalable than full 6-DoF motion estimation, but only provide a coarse localization (that of the closest database image). Therefore, place recognition methods are often used as an initial step to constrain the 6-DoF localization to a few images.

Classical visual place representations consist of handcrafted local descriptors aggregated with either Bag-of-Words (BoW) [10], Fischer vectors [11] or Vectors of Locally Aggregated Descriptors (VLAD) [12]. Learning place representations using deep networks has boosted the capabilities and performance of place recognition. The architectures consist of a convolutional backbone followed by a pooling operation, such as max-pooling [13] or average-pooling [14]. Radenovid *et al.* [15] proposed a Generalized Mean Layer (GeM) that learns the norm of the pooling-operator, and thus generalizes max- and average-pooling. Arandjelovic *et al.* [16] proposed NetVLAD, a deep architecture that also learns the VLAD clusters. MultiViewNet [17] and Warburg *et al.* [5] incorporate multiple views to improve retrieval performance. The Bayesian triplet loss [18] mirrors the triplet loss, but allows a network to embed images into Gaussian distributions rather than points, and thus propagate uncertainties to image retrieval. More similar to our work, Sourav *et al.* [19] explored extreme viewpoint changes by having query and database images from opposite directions.

B. 6-DoF Visual Localization

Methods for camera localization have traditionally been classified as either structure-based or regression-based [20]. **Regression-based** methods train a deep network to directly regress the camera pose from an input image. Some notable approaches are PoseNet [21], that estimates the absolute pose of a camera with respect to a scene, and the works by Laskar *et al.* [22] and Balntas *et al.* [23], that estimate the relative pose between two cameras. However, recent evaluations (Zhou *et al.* [24] among others) seems to show that direct pose regression is less accurate than the more traditional one based on feature extraction and matching.

Structure-based methods, on the other hand, predict the pose of the camera by matching features between a 3D model and 2D query images. Traditional handcrafted descriptors struggle to match images taken under strongly differing viewing conditions. [25] tackle large variations in viewing angle by rendering SIFT features from views with less extreme view-angle difference. Modern localization methods rely on convolutional neural networks to extract features that are more robust to appearance and viewpoint changes. SuperPoint [26] consists of a convolutional encoder followed by two heads: one for classifying if a pixel is an interest point, and the other to encode a feature descriptor. D2-Net [27] has a single CNN that extracts dense features that serves both as descriptors and detectors. LOFTR [28], on the other hand, takes a pair of images as input and via a ViT [29]-based transformer architecture estimates both keypoints and matches simultaneously. Another line of research has focused on learning local descriptors using image level supervision only [30], [31], [32], [33]. DELF [30] learns a spatial attention that is used to pool the feature map and can thus be optimized similarly to retrieval networks, but via the attention mechanism yields local features. Combining networks that predict both a coarse place descriptor and local descriptors [34], [35], [36], [37], [38] have shown to improve both efficiency and robustness.

C. Visual Localization Datasets

Many large localization datasets have been proposed in recent years, focusing mainly on viewpoint and appearance changes. Among the most relevant **place recognition datasets** we can cite Nordland [3] with seasonal changes, Tokyo24/7 [4] with day-night changes, and MSLS [5], which is currently the largest and most diverse place recognition dataset including viewpoint, structural, seasonal and day-night changes. **6-DoF datasets** include ground truth poses that are typically obtained from SfM reconstructions or differential GPS. Oxford Robotcar [6] traverses the same loop 100 times during a year in varying weather and day/night conditions. Extended CMU Seasons dataset [7], [39] is similarly recorded with a car-mounted camera. Aachen Day-Night [7], [40] contains images from hand-held devices and focuses on day-night changes. The recent ETH-Microsoft [8] covers indoor environments and challenging day/night appearance changes. All these datasets only contain street-level images from ground vehicles or handheld devices. In contrast, our DAG dataset contains images taken by a ground vehicle and an airplane.

D. Aerial-to-Street-Level Retrieval and Localization

Aerial-to-Street-Level registration was addressed by Shan *et al.* [41]. Similarly to us, they propose a view-dependent matching process. However, they assume to know the approximate street-level position from GPS EXIF tags, while

Fig. 2: The airplane has five cameras mounted facing downwards, West, South, North and East. The figure shows images from the suburban environment in the *Lolland* sequence and the harbor environment at the *Nordhavn* traversal.

we run a deep place recognition model to obtain such coarse localization. As a second difference, their 3D reconstructions are created from the street level, which is only possible when multiple ground-level images of the same area are available. We show that accurate 3D models and street-level renders can be estimated from aerial images This generalizes easily to scenes with few street-level data.

Lin *et al.* [42] propose to train a place recognition network for direct aerial-to-street-level retrieval. They construct a dataset that covers several large cities with both aerial and street level images. They train a place recognition network to be invariant to the extreme viewpoint change between aerial and street level images. In contrast to their work, we seek to find local correspondences to improve the coarse localization estimate of the place recognition model.

Reducing viewpoint differences between aerial and ground images using generative adversarial networks (GANs) was explored in [43], [44], [45], [46]. These methods are valid alternatives to our baseline. Note, however, that their view syntheses come with no guarantees and might produce low-level artifacts that affect the localization performance.

[47] is the most similar to our work, releasing a 2 kilometers-long sequence captured by a drone in Zürich and Google Maps images from the same places. Our DAG dataset is significantly larger, covering over 50 kilometers in more diverse urban and suburban environments. Their ground-truth car poses come from GPS and have limited accuracy. In contrast, our street-level images are annotated with differential GPS and are thus more accurate.

III. The Danish Airs and Grounds (DAG) Dataset

DAG contains aerial and street-level images from urban, suburban and rural regions in Denmark[1]. The airplane photos are taken by five cameras; one facing vertically downward, and four oblique views facing each of the world corners (East, West, North, South). The images were recorded by The Danish Agency for Data Supply and Efficiency in 2017 and 2019. See Fig. 2 for examples or visit their website for an interactive look at the images[2]. The airplane photos are recorded at an 150 meter altitude and with a spatial frequency between $20-70$ meter. Their associated poses are in principle within 5 meter precision, which is further improved by visual alignment and multi-view reconstruction.

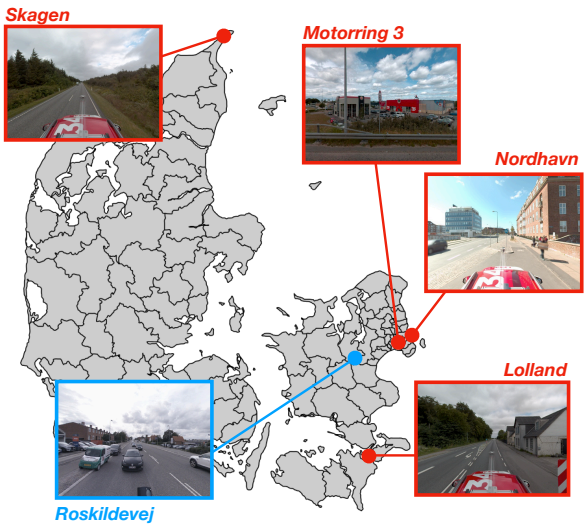


Fig. 3: The five DAG sequences cover a large geographical area in Denmark. The sequences are captured at urban, suburban and rural regions, as highlighted with an example image from each sequence. The four red sequences are used for the training set and the blue sequence kept as the test set.

The street-level images are recorded with a Ladybug5+ by the Danish Road Directory. Fig. 3 shows the location of the five sequences; the *Nordhavn* sequences consist of three sequences recorded in an urban harbor environment from a boat and a car. The *Motorring 3* sequence is a suburban road around Copenhagen, the *Roskildevej* sequence is in a urban environment, and both the *Skagen* and *Lolland* sequences are from rural areas in Denmark. The recorded sequences cover more than 50 km and have more than $11,000$ panoramic images, which we project into four perspective

[1]The access to the data was possible thanks to the open access policy of the Danish Government, c.f. https://dataforsyningen.dk/.

[2]https://skraafoto.kortforsyningen.dk

Fig. 4: Street-level-to-aerial localization pipeline. We generate a dense 3D reconstruction from the aerial images. From this reconstruction we render a database of photo-realistic images for real-image-to-render retrieval from street-view queries.

cameras, totaling over $44,000$ images. Fig. 3 shows some examples of street-level images from different environments. The trajectories and the photo orientation is determined by a combination of differential GNSS and IMU. In particular the NovaTel IMU-FSAS [3] is used, which measures roll, pitch, heading with high precision. Following, the trajectories are refined using signaled fixed points to achieve a precision of approximately 5 cm.

IV. Street-level-to-Aerial Localization

We implemented a localization pipeline that can be denoted as image-to-render, an intermediate category between image-to-image and image-to-map matching (following the terminology of [48]). The extreme parallax angle and scale change between aerial and ground-level images renders image-to-image matching very challenging. Estimating intermediate 3D representations and rendering synthetic images at ground-level allows us to bridge viewpoint challenges and leverage recent deep models for image-to-image matching. Our experiments show that the appearance differences between real and rendered images are not an issue when matching deep features.

Our method consists of the following steps, which are also depicted in Fig. 4: *First*, we create a 3D model from the aerial images. *Second*, we render street-level images from this 3D model in a regular grid. *Third*, we use a place recognition method to retrieve street-level renderings from the same place as a given street-level query image. *Fourth*, we use a structure-based localization method between the retrieved rendered image and the query image to obtain the 6-DoF pose of the query image.

A. 3D Reconstruction and Ground-Level View Synthesis

We used the commercial software Agisoft Metashape[4] to generate dense 3D reconstructions from the aerial images. Due to the large computational and memory footprint of the 3D models, we partitioned each sequence into sub-models of approximately 2 kilometers. After that, we synthesize ground views of the 3D model in a regular grid with 5-meter separation between synthetic cameras. We synthesize 8 street-level renderings at each location at equally spaced directions ($45°$ between each other). We set the intrinsics of the synthetic perspective cameras as the same as the camera used to record the query street-level images.

B. Place recognition

We use a Resnet50 followed by the GeM aggregation layer [15] as our place recognition network. We trained the network with the triplet loss and hard negative mining. We found that pre-training on the MSLS [5] significantly improves the retrieval performance. Fig. 5 shows examples of some of the triplets presented to the network during training. Note that the anchor and the positive are from the same place and the negative is from a different place. We found experimentally that it is important that the anchor and the positive image in each triplet are of the same type, either both synthetic or both real images.

C. 6-DoF Visual Localization

Once the initial place is retrieved, our method proceeds with the actual 6-DoF localization, which is based on a Perspective-n-Point (PnP) solver [49]. The goal is to find the camera pose that, given a set of 3D points, minimizes the reprojection error of the 2D points in the camera plane. The peculiarity in our case is that the 3D points are calculated by back-projecting the pixels of the rendered camera with associated depth information, while the 2D projections are extracted from the original picture's corresponding pixels.

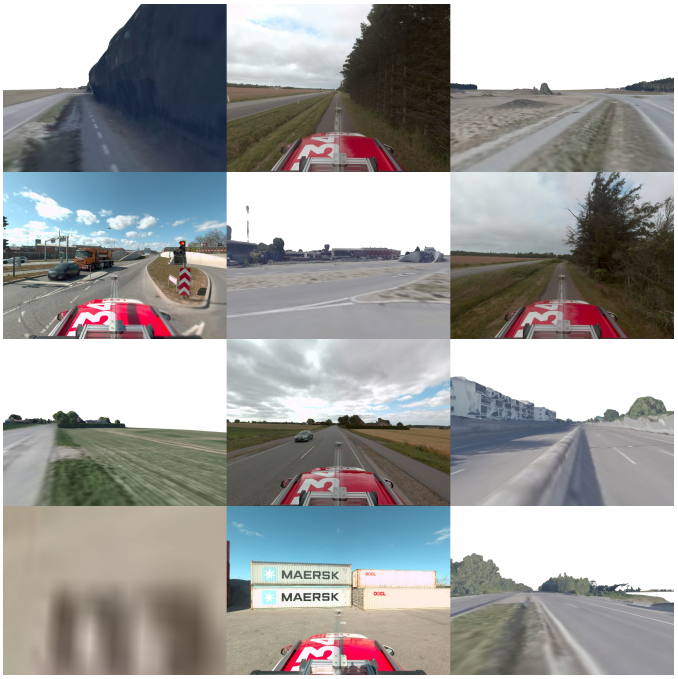
We experimented with both SIFT [50] and D2-Net [27] as feature detectors and descriptors. We use the ratio test [50] to filter matches for SIFT, but use a cross-matching check for D2-Net as suggested by the authors [27]. We use only the 1000 best matches to increase the chances for RANSAC convergence. Once the rendered picture matches were identified, each pixel in the rendered image was backprojected to obtain its 3D coordinates in the world using the depth of the 3D reconstruction. We then use a PnP solver to obtain the 6-DoF pose between the 2D and 3D point correspondences.

[3]https://hexagondownloads.blob.core.windows.net/public/Novatel/assets/Documents/Papers/FSAS/FSAS.pdf

[4]https://www.agisoft.com/

	R@1	R@5	R@10	R@20	M@5	M@10	M@20
Triplet R50 (DAG)	0.68	0.78	0.82	0.87	0.63	0.60	0.56
Triplet R50 (MSLS)	0.35	0.48	0.55	0.64	0.29	0.27	0.24
Triplet R50 (MSLS + DAG)	**0.80**	**0.90**	**0.92**	**0.95**	**0.77**	**0.74**	**0.71**

TABLE I: Recall (R) and Mean Average Precision (M) at $\{1, 5, 10, 20\}$. Both methods consist of a ResNet50 followed by a Generalized Mean (GeM) layer. Pre-training on MSLS significantly improves the coarse localization performance.



Anchor Positive Negative

Fig. 5: Examples of training triplets. Note that in the harbor environment, *Nordhavn*, the containers are not in the same location when the car and the airplane visited the site. As seen in the last row, the rendering that is geographically close to the car, shows just the front of a container. These structural changes make visual localization challenging.

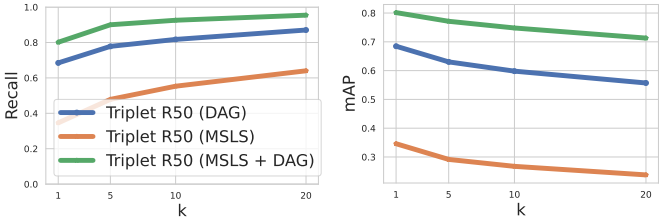


Fig. 6: Recall and mean average precision (mAP) at k for a Resnet50 with GeM pooling trained with the triplet loss on DAG, on MSLS, and finally pre-trained on MSLS and fine-tuned on DAG (MSLS+DAG).

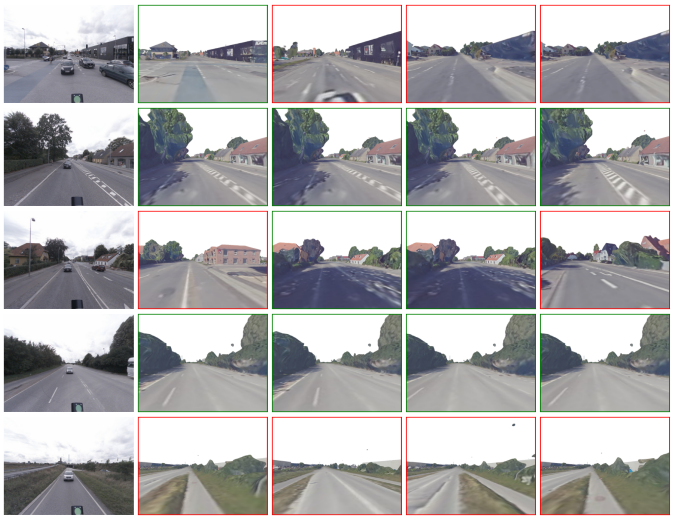


Fig. 7: Qualitative retrieval results. 1^{st} column: query images. 2^{nd} to 5^{th} columns: in order, four closest renders retrieved from the database. Green/red boundaries indicate distances between query and retrievals smaller/greater than 25 meters. Observe that the network learned invariances to textural differences between rendered and real images. Our method is especially challenged in areas without buildings.

V. Experimental Results

A. Place Recognition

Fig. 6 and Table I show the recall@k and the mean average precision (mAP@k), evaluated at k number of nearest neighbors, on the test sequence *Roskildevej*. We use the standard threshold of 25 meters for true positives. A Resnet50 with a GeM-layer, trained with the triplet loss (Triplet R50 (DAG)) correctly retrieves the same-place database image 68% of the times (Recall@1 is 0.68) in the test set. Pre-training the network on the very large place recognition dataset MSLS [5], and then fine-tuning on DAG, results in a significantly improved performance (Triplet R50 (MSLS + DAG)). With this setup, the Recall@1 increases to 0.80. The model pre-trained at MSLS without fine-tuning in our DAG data (Triplet R50 (MSLS)) performs significantly worse than the first two, due to the domain gap

Fig. 7 shows some qualitative examples of the network retrievals. The network struggles in scenes with dynamic objects and vegetation. We believe that vegetation is a particularly challenging instance of this dataset. The aerial images and street-level images are not taken at the same time, thus trees and bushes change appearance (summer/winter). Furthermore, one of the limitations of our localization pipeline is that the 3D reconstruction of vegetation is very coarse. As the aerial images are not taken at the same time, changes in the vegetation (motion caused by the wind, vegetation growth or seasonal effects) result in a smoothing of the 3D reconstruction. Research into direct aerial to street-level localization (without 3D reconstruction) or learning methods that consider such changes are promising directions as they

can circumvent this limitation.

B. Visual Localization

In this section we evaluate the localization error of the relative pose between a query and the rendered image retrieved as described in Section IV-B. We evaluate several alternatives: 1) assigning to the query the pose of the closest database image (denoted as NN), 2) using SIFT/D2-Net features and PnP as described in Section IV-C.

Fig. 8 shows cumulative error plots (fraction of images with translational and rotational errors under different thresholds). SIFT offers the worst results, due to the differences between low-level textures in real images and rendered ones. D2-Net performs significantly better, as it learned higher level, more semantically meaningful description of the features, that is less dependent on specific low-level texture patterns. Our best median errors are around 5 meters and 7 degrees, which is remarkable given the challenging nature of the data.

Note finally that D2-Net matches outperform the rotation estimates of a simple nearest-neighbour matching, but not in translation. We attribute this to the aggregation of the errors involved in the visual localization estimate. The aerial image resolution is 10 centimeters per pixel. Assuming matching errors over 1 pixel, parallax angles between 20° and 40° and small translation and rotation errors, they propagate to triangulation errors over 1 meter. Such reconstruction errors may be bigger for textureless areas, vegetation and dynamic objects, and propagate to the localization via PnP. Our optimal RANSAC threshold is 40 pixels, which indicates that there exist matches with high error that also add up to the localization error. We also observed unevenly distributed matches. Simulations of the geometry of the problem gave errors of the same level as those obtained with the real data.

Fig. 9 shows several examples of D2-Net matches between the query and database images after cross matching. Observe how the features extracted on buildings have in general low image errors. Matches on the road and in vegetation, on the other hand, have a coarser localization in the image.

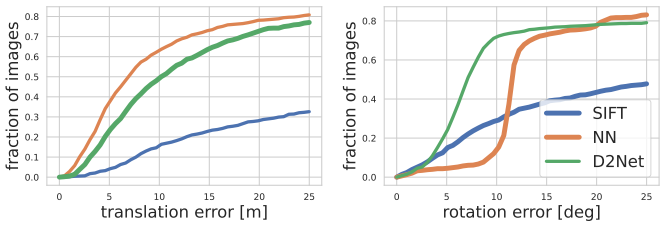


Fig. 8: Cumulative translation and rotation errors for nearest-neighbour (NN), SIFT-based and D2-Net-based pose.

	5m/5°	10m/10°	25m/25°
Triplet R50 (MSLS + DAG) + NN	0.02	0.10	**0.75**
Triplet R50 (MSLS + DAG) + SIFT	0.01	0.11	0.29
Triplet R50 (MSLS + DAG) + D2-Net	**0.04**	**0.42**	0.66

TABLE II: Ratio of queries under several pose thresholds.

Table II shows the ratio over *all* retrievals of those with error under certain translation/rotation errors (e.g., 42% of all queries have pose errors below 10m/10°). Observe how D2-Net features and PnP improve over the retrieval poses for well conditioned images (thresholds 5m/5° and 10m/10°). However, pose estimates in cases with noisy matches and reconstructions increase the error with respect to retrieval (threshold 25m/25°). This suggests that further research is needed for wide baseline feature matching, and we hope our DAG dataset is useful for this task.

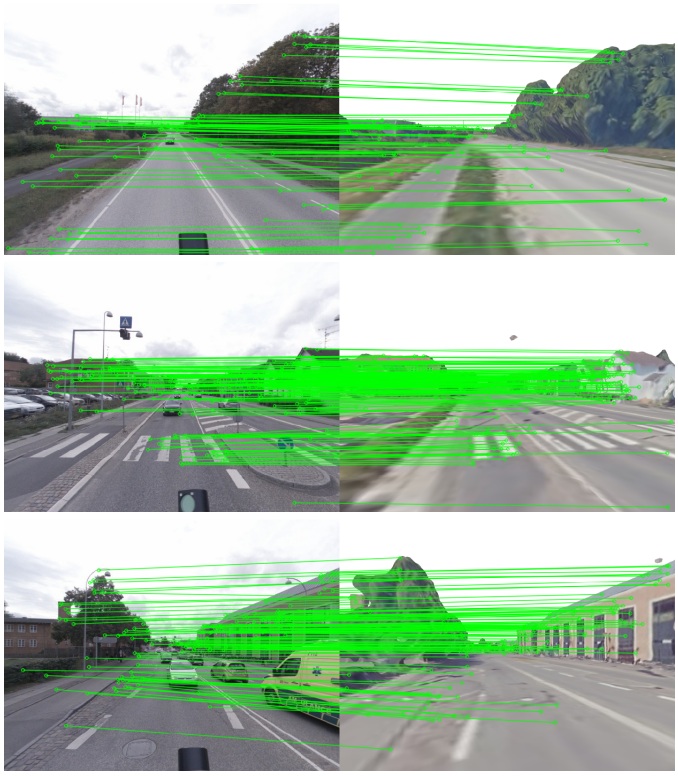


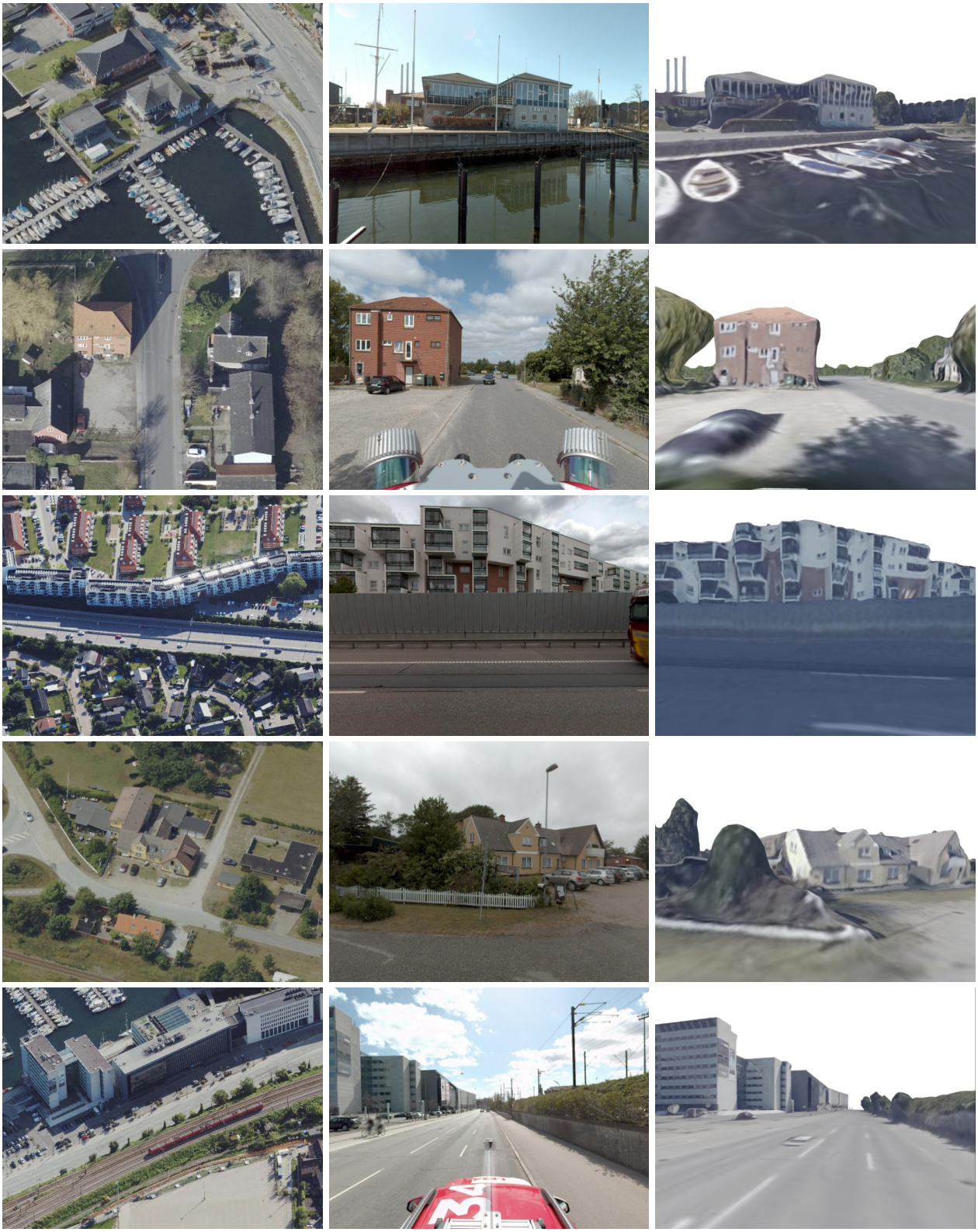
Fig. 9: Inlier matches between query images and the street-level renderings. Since aerial images are taken at different times, moving objects such as cars and leaves are smoothed out or blurred. However, D2-Net is still able to find reliable matches between the two views. The challenging aerial-to-street-level viewpoint changes are alleviated via 3D renderings and D2-Net produces reasonable results.

VI. Additional Visualizations of the Data

In Figure 10, we present additional visualizations from the dataset. These images highlight again the difficulty of the problem and the diversity of the DAG dataset, covering urban, suburban and rural areas. As seen in the second row, the dataset also includes seasonal and dynamic changes between the aerial and street level image.

VII. Conclusions

In this paper, we have presented *Danish Airs and Grounds* (DAG), a dataset for aerial to street-level visual localization. Our data collection is the largest, up to date, that addresses such challenging setup. We believe there are two main aspects that make DAG relevant for the robotics and computer vision



(a) Aerial Images (b) Street-View Images (c) Street-View Renderings

Fig. 10: Sample visualizations of aerial, street-level images and street level renderings. The images illustrate the diversity of scenes in our DAG dataset.

communities. Firstly, it addresses a particular case of wide baseline matching, which is one of the hardest cases for retrieval and localization. And secondly, from a more practical perspective, targets the relevant application case of street-level localization in aerial maps. We present a simple map-to-image re-localization pipeline for wide-baseline matching. In our experiments we analyze the performance of such an approach, serving as validation and initial baseline for our dataset.

REFERENCES

[1] M. Cummins and P. 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.

[2] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[3] NRK. (2013) Nordlandsbanen: minute by minute, season by season. [Online]. Available: https://nrkbeta.no/2013/01/15/nordlandsbanen-minute-by-minute-season-by-season/

[4] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *IEEE/CVF CVPR*, 2015.

[5] F. Warburg, S. Hauberg, M. Lopez-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," in *IEEE/CVF CVPR*, 2020.

[6] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[7] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla, "Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions," in *IEEE/CVF CVPR*, 2018.

[8] ETH Zurich Computer Vision Group and Microsoft Mixed Reality & AI Lab Zurich, "The ETH-Microsoft Localization Dataset," https://github.com/cvg/visloc-iccv2021, 2021.

[9] Z. Zhang, T. Sattler, and D. Scaramuzza, "Reference pose generation for long-term visual localization via learned features and view synthesis," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 821–844, 2021.

[10] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *IEEE/CVF ICCV*, 2003.

[11] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed Fisher vectors," in *IEEE/CVF CVPR*, 2010.

[12] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *IEEE/CVF CVPR*, 2010.

[13] H. J. Giorgos Tolias, Ronan Sicre, "Particular object retrieval with integral max-pooling of cnn activations," *ICLR*, 2016.

[14] A. Babenko and V. S. Lempitsky, "Aggregating deep convolutional features for image retrieval," *IEEE/CVF ICCV*, 2015.

[15] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[16] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *IEEE/CVF CVPR*, 2016.

[17] J. M. Facil, D. Olid, L. Montesano, and J. Civera, "Condition-invariant multi-view place recognition," *arXiv preprint arXiv:1902.09516*, 2019.

[18] F. Warburg, M. Jørgensen, J. Civera, and S. Hauberg, "Bayesian triplet loss: Uncertainty quantification in image retrieval," *IEEE/CVF ICCV*, 2021.

[19] S. Garg, N. Suenderhauf, and M. Milford, "Semantic–geometric visual place recognition: a new perspective for reconciling opposing views," *The International Journal of Robotics Research*, p. 0278364919839761, 2019.

[20] A. Zhou, "Survey on visual-based localization."

[21] A. Kendall, M. Grimes, and R. Cipolla, "Convolutional networks for real-time 6-dof camera relocalization," *IEEE/CVF ICCV*, 2015.

[22] Z. Laskar, I. Melekhov, S. Kalia, and J. Kannala, "Camera relocalization by computing pairwise relative poses using convolutional neural network," *IEEE/CVF CVPRW*, 2017.

[23] V. Balntas, S. Li, and V. Prisacariu, "Relocnet: Continuous metric learning relocalisation using neural nets," in *ECCV*, 2018.

[24] Q. Zhou, T. Sattler, M. Pollefeys, and L. Leal-Taixe, "To learn or not to learn: Visual localization from essential matrices," in *IEEE ICRA*, 2020.

[25] D. Sibbing, T. Sattler, B. Leibe, and L. Kobbelt, "Sift-realistic rendering," in *3DV*, 2013.

[26] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," *IEEE/CVF CVPRW*, 2017.

[27] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable CNN for joint detection and description of local features," *IEEE/CVF CVPR*, 2019.

[28] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," *IEEE/CVF CVPR*, 2021.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[30] H. Noh, A. Araujo, J. Sim, and B. Han, "Image retrieval with deep local features and attention-based keypoints," *IEEE/CVF ICCV*, 2016.

[31] G. Tolias, T. Jenícek, and O. Chum, "Learning and aggregating deep local descriptors for instance-level recognition," *ECCV*, 2020.

[32] I. Rocco, M. Cimpoi, R. Arandjelovic, A. Torii, T. Pajdla, and J. Sivic, "Neighbourhood consensus networks," *NeurIPS*, 2018.

[33] G. Kurzejamski, J. Komorowski, L. Dabala, K. Czarnota, S. Lynen, and T. Trzcinski, "Superncn: Neighbourhood consensus network for robust outdoor scenes matching," in *ACIVS*, 2020.

[34] P. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," 2018.

[35] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *ECCV*, 2020.

[36] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," *IEEE/CVF CVPR*, 2021.

[37] T.-Y. Yang, D.-K. Nguyen, H. Heijnen, and V. Balntas, "Ur2kid: Unifying retrieval, keypoint detection, and keypoint description without local correspondence supervision," *arXiv preprint arXiv:2001.07252*, 2020.

[38] P. Weinzaepfel, T. Lucas, D. Larlus, and Y. Kalantidis, "Learning super-features for image retrieval," *arXiv preprint arXiv:2201.13182*, 2022.

[39] H. Badino, D. Huber, and T. Kanade, "The CMU Visual Localization Data Set," http://3dvis.ri.cmu.edu/data-sets/localization, 2011.

[40] T. Sattler, T. Weyand, B. Leibe, and L. P. Kobbelt, "Image retrieval for image-based localization revisited," in *BMVC*, 2012.

[41] Q. Shan, C. Wu, B. Curless, Y. Furukawa, C. Hernandez, and S. M. Seitz, "Accurate geo-registration by ground-to-aerial image matching," in *3DV*, 2014.

[42] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *IEEE/CVF CVPR*, 2015, pp. 5007–5015.

[43] X. Lu, Z. Li, Z. Cui, M. R. Oswald, M. Pollefeys, and R. Qin, "Geometry-aware satellite-to-ground image synthesis for urban areas," in *IEEE/CVF CVPR*, 2020.

[44] K. Regmi and M. Shah, "Bridging the domain gap for ground-to-aerial image matching," in *IEEE/CVF ICCV*, 2019.

[45] K. Regmi and A. Borji, "Cross-view image synthesis using geometry-guided conditional gans," *Computer Vision and Image Understanding*, vol. 187, p. 102788, 2019.

[46] ——, "Cross-view image synthesis using conditional gans," in *IEEE/CVF CVPR*, 2018.

[47] A. L. Majdik, D. Verda, Y. Albers-Schoenberg, and D. Scaramuzza, "Air-ground matching: Appearance-based GPS-denied urban localization of micro aerial vehicles," *Journal of Field Robotics*, vol. 32, no. 7, pp. 1015–1039, 2015.

[48] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, and J. Tardós, "An image-to-map loop closing method for monocular slam," in *IEEE/RSJ IROS*, 2008.

[49] G. Terzakis and M. Lourakis, "A consistently fast and globally optimal solution to the perspective-n-point problem," in *ECCV*, 2020.

[50] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.