CrossLocate: Cross-modal Large-scale Visual Geo-Localization in Natural Environments using Rendered Modalities

Jan Tomešek, Martin Čadík, Jan Brejcha
Brno University of Technology, Faculty of Information Technology, CPhoto@FIT
Božetěchova 2, 61200 Brno, Czech Republic
{itomesek,cadik,ibrejcha}@fit.vutbr.cz

Abstract

*We propose a novel approach to visual geo-localization in natural environments. This is a challenging problem due to vast localization areas, the variable appearance of outdoor environments and the scarcity of available data. In order to make the research of new approaches possible, we first create two databases containing "synthetic" images of various modalities. These image modalities are rendered from a 3D terrain model and include semantic segmentations, silhouette maps and depth maps. By combining the rendered database views with existing datasets of photographs (used as "queries" to be localized), we create a unique benchmark for visual geo-localization in natural environments, which contains correspondences between query photographs and rendered database imagery. The distinct ability to match photographs to synthetically rendered databases defines our task as "cross-modal". On top of this benchmark, we provide thorough ablation studies analysing the localization potential of the database image modalities. We reveal the depth information as the best choice for outdoor localization. Finally, based on our observations, we carefully develop a fully-automatic method for large-scale cross-modal localization using image retrieval. We demonstrate its localization performance outdoors in the entire state of Switzerland. Our method reveals a large gap between operating within a single image domain (*e.g. *photographs) and working across domains (*e.g. *photographs matched to rendered images), as gained knowledge is not transferable between the two. Moreover, we show that modern localization methods fail when applied to such a cross-modal task and that our method achieves significantly better results than state-of-the-art approaches. The datasets, code and trained models are available on the project website:* `http://cphoto.fit.vutbr.cz/crosslocate/`.

1. Introduction

Visual geo-localization aims to estimate the geographical origin of a visual document, *i.e.* a photograph or a video.

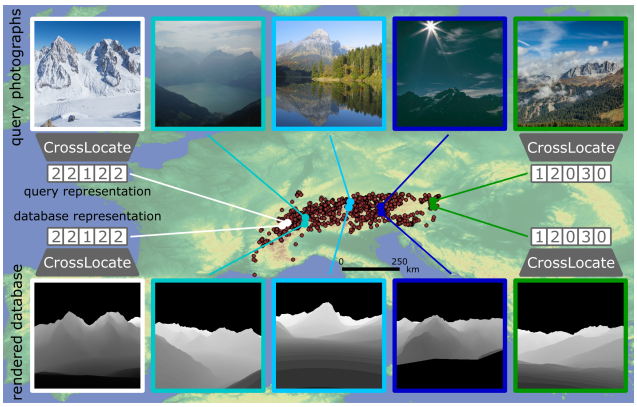

Figure 1. **CrossLocate** localizes ground-level photographs captured in diverse environments across the Alps. Our new databases of rendered image modalities enable the implementation of cross-modal image-retrieval (*i.e.* matching real photographs to rendered imagery). Out of several assessed modalities, we select depth maps and train a cross-modal global descriptor for large-scale image localization outdoors.

The ability to localize (ground-level) photographs enables many applications including autonomous navigation, augmented reality, advanced image enhancements, automatic organization of photo collections, and other historical and forensic tasks, where geographical information has been lost or intentionally removed. The ultimate goal is to accurately determine the position and orientation of an image captured anywhere in the world. Unfortunately, this is far from possible with current methods. The complexity of the task leads to the fact that localization methods need to focus on a specific environment and specific scale.

Visual localization has been researched for several decades [21, 44, 3]. While tremendous progress could be observed in the localization focused on outdoor urban areas [9, 17], localization targeted in nature remains an open problem. Existing methods localize with an error in order of kilometers at best [35, 6], or assume a good initial pose estimate and only attempt to refine the camera pose [7, 33, 12].

Natural environments introduce a number of specific challenges. They are highly variable, as both their appearance and geometry change with weather and seasons [5]. The amount of available data (*i.e*. user-taken photographs), is low compared to urban areas, and the spatial coverage is sparse and uneven. Accordingly, it is often impossible to localize in nature using only real photographs. Existing localization approaches, therefore, operate across multiple domains or image modalities, such as cross-view methods [25, 29, 30] utilizing satellite and aerial imagery or cross-modal methods utilizing 3D terrain models [35].

We focus on visual geo-localization of ground-level photographs captured in large natural areas, such as the Alps, without the use of additional sensors. We utilize a digital elevation model for building databases of rendered views to be later matched with query photographs (a "cross-modal" setup) (Fig. 1). We claim the following **contributions**:

- We create two unique databases by rendering image modalities (semantic segmentations, silhouette maps, depth maps) from the entire Alps. We combine these databases with query photographs to form a new benchmark for localization in natural environments.
- We provide an insight into the localization performance of the image modalities, and suggest depth information as the best choice.
- We carefully design a cross-modal deep learning method for localization within mountains (the Alps), based on a weak localization supervision. Comparison with previous work shows that our method achieves state-of-the-art results.
- We demonstrate the difficulty when moving from a single-domain approach (*i.e*. photographs) to the processing of multiple domains (*i.e*. photographs and renders). The behavior observed within a single domain is not directly transferable to multiple domains, and there is only a small benefit to synthetic pre-training.

2. Related work

Two significant aspects that influence the design of localization methods are target environment (*urban* [3, 43, 27, 10, 40, 4], *natural* [35, 12, 7, 33], *global* [21, 22, 44, 41]) and spatial scale (*city-scale* [3, 43, 30], *large-scale* [35], *planet-scale* [21, 22, 44, 41]). The vast differences between the individual environments and scales lead to diverse approaches. As a result, many underlying localization principles may be observed (classification [44, 19, 4], retrieval [3, 35, 21, 22, 41, 40, 34, 32], regression [27, 9, 10], structure-from-motion [20, 1, 24, 18]).

Global planet-scale approaches [21, 22, 44, 41] attempt to localize images captured anywhere in the world, no matter the environment, which usually leads to localization errors in hundreds of kilometers. Therefore, these methods may be useful for space pruning and scene type recognition. Specifically, PlaNet [44] is a deep learning classification approach to geo-localization. The classification approach was later shown inferior to the image retrieval utilized by Revisited IM2GPS [21, 41]. This highlights the advantage of building a general image descriptor in comparison to trying to memorize the entire world within a classification model. Furthermore, the retrieval approach required less data while providing better performance. This is an important observation with respect to natural environments where data is extremely scarce.

Approaches aimed at outdoor *(sub)urban* environments [40, 4] are much more advanced and precise, as they have gained a lot of attention in recent years. They are typically used for *city-scale* localization [3, 43, 30], though some are precisely tuned for specific places or landmarks [9, 27]. While these *city-scale* approaches were designed for urban areas, they might represent a potential avenue to the solution of localization in nature. NetVLAD [3] successfully utilizes the retrieval approach. It combines custom feature aggregation with weakly supervised learning to perform place recognition despite changes in appearance over time. Unfortunately, the NetVLAD aggregation results in large descriptors, which are not ideal for our large-scale localization, even after the proposed dimensionality reduction to 4096 dimensions. However, a further reduction at the cost of accuracy might be possible. HOW [39] is the state-of-the-art instance-level recognition (retrieval/localization) method trained on datasets of outdoor photographs of landmarks and buildings. It uses learned internal local descriptors (HOW) combined with an ASMK image search [38] approach to perform search and classification in the domain of landmarks, where it outperforms existing global and local descriptors. To a certain extent, the ASMK is considered a replacement of traditional spatial verification. DELG [14] is another state-of-the-art large-scale image retrieval approach. Contrary to HOW, it utilizes both the global and local descriptors learned within a single model to perform two-step retrieval and instance-level recognition for outdoor landmark scenes.

Regression approaches to urban localization [9, 27] can provide sub-meter spatial precision, but they are usable only on small areas. Urban localization is often solved through structure-from-motion (SfM) [20, 1, 24]. SfM approaches [26, 46, 18] perform localization using 3D models acquired from many overlapping photographs. This requires millions of photographs, which makes SfM approaches unsuitable for large-scale localization in nature.

Methods focusing specifically on *natural* environments are far less explored. They often operate at a *large scale* [35], corresponding to an area of a country or mountain range. A separate group of methods focuses only on camera pose refinement from a good initial estimate [7, 33]. *Global* and *urban* approaches typically utilize only ordinary

photographs for localization. However, in *natural* environments, the number of user-taken photographs is low and the spatial coverage is sparse and uneven. This leads to various methods of cross-view or cross-modal character. The cross-view methods [25, 29, 30, 42, 31, 37] utilize databases of satellite or aerial imagery for the localization of ground-level queries. The cross-modal methods make use of digital elevation models to create databases of various synthetic modalities, such as skylines [35]. Synthetic modalities (*e.g.* horizon lines and silhouette maps) were successfully used for a camera pose estimation in nature [7, 33]. Semantic segmentations were used for camera pose estimation both in nature [12, 8] and in a city [4]. Assuming that localization methods are powerful enough to work across image domains, synthetic modalities might offer a solution to the localization in nature.

Close to our work, horizon-based localization [6, 35] (abbreviated HLoc) localizes photographs captured in mountains by extracting visible skylines and comparing them with synthetic horizon curves stored in a database. While skylines arguably carry useful information, we show that other features might be more efficient, especially in situations where the horizon is obscured or not visible.

3. CrossLocate datasets

Existing datasets of photographs captured in nature [35, 11, 13] are rare and provide photographs usable only as queries, as they are typically few in number (hundreds or low thousands) and sparsely distributed. To enable the development of novel visual geo-localization methods in natural environments, and to complement the existing query datasets, we created two novel databases of "synthetic" ground-level views – spatially non-uniform *"sparse"* database and *"uniform"* database. Each view is accompanied by detailed information about its position and orientation. The *sparse database* serves for fast and simple experiments, while the *uniform database* represents the real-world scenario of localization across a large area of hundreds of thousands of square kilometers and millions of images (see Fig. 2). Each of these databases contains three rendered image modalities – *semantic segmentations*, *silhouette maps* and *(absolute) depth maps*. Other modalities can be derived, such as previously used horizon lines or relative depth maps, thus enabling diverse tasks.

In our terminology, a "dataset" consists of "queries" to be localized and a "database" to be searched. Detailed dataset information and the procedure behind the creation of the image modalities are presented in the supplementary material (Sec. 2).

3.1. Query photographs

In order to avoid a need for the creation of custom datasets of query photographs, and to design our databases fittingly, we first choose two existing datasets of photographs captured in the region of the Alps – the GeoPose3K dataset [11] and the Landscape AR dataset [13].

The GeoPose3K dataset consists of 3111 photographs with a manually verified position and orientation of each image. We use all the images as queries to be localized.

The Landscape AR dataset consists of 16K photographs automatically collected from the Internet, with their positions and orientations estimated using Structure-from-Motion. We use 9K of these photographs and combine them with the GeoPose3K dataset, thus creating a "CrossLocate" query dataset of 12353 photographs from across the Alps.

We introduce geographically disjoint splits into training, validation and testing sets, where the area of Switzerland (40,000 km^2) is put aside for testing. The same testing area was also used by HLoc [35].

3.2. Sparse dataset

The *sparse database* is aimed at small, fast and simple experimentation. We use all the 3111 query positions from the GeoPose3K dataset and render synthetic views at these positions. At each of these positions, we render 12 views of resolution 500×500 pixels, each with a 60° field of view. There is a 30° difference in the yaw angle of the individual views, thus covering the whole 360° view range. This results in 37332 views forming the *sparse database*. In this way, the *sparse database* represents the smallest possible database created without sacrificing any of the 3111 valueable query photographs.

Since the query positions correspond to the database positions, the contents of query photographs and corresponding database views are of a similar scale. Therefore, the localization task is simplified. There is, however, an expected difference in pitch, roll and especially yaw angle, which leads to natural misalignment between the queries and the database, which is left for localization methods to solve.

We also propose a split into training, validation and testing sets, where the area of Switzerland is assigned to the testing set. Tab. 1 shows the pairing of the *sparse database* with the query photographs from the GeoPose3K dataset, and the resulting splits. We designate this as the *Sparse dataset*.

Sparse dataset	train	val	test	total
sparse database	16908	6192	14232	**37332**
queries (GeoPose3K [11])	1409	516	1186	**3111**

Table 1. Numbers of query and database images in the *Sparse dataset*, and their split into training, validation and testing sets.

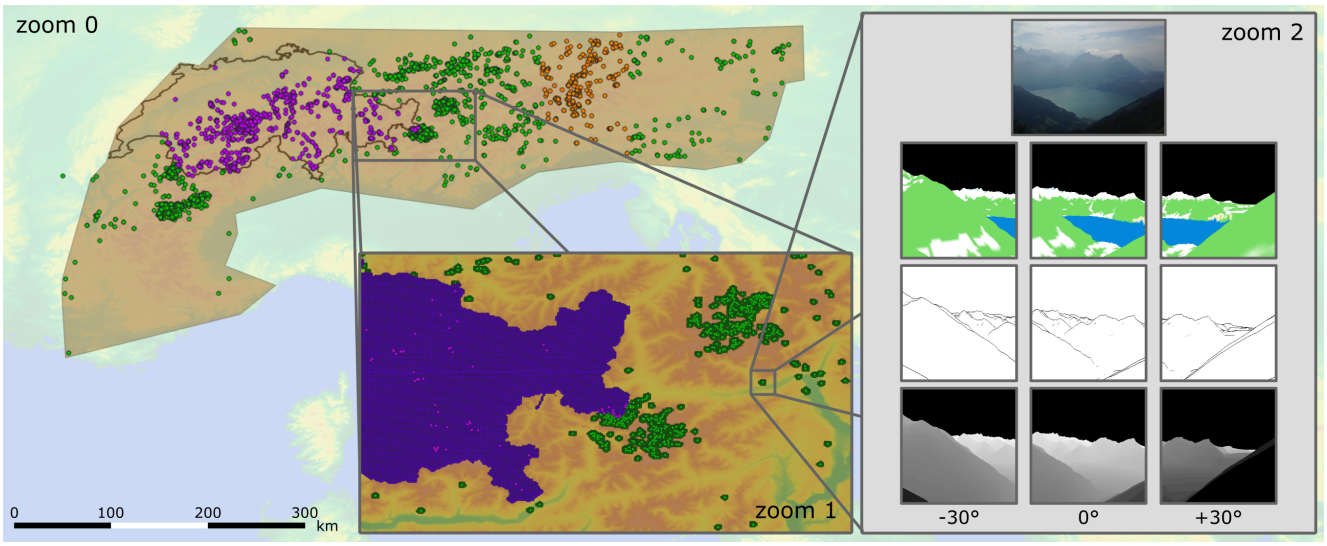

Figure 2. **Our datasets** at various zoom levels. **Zoom 0** shows positions of query photographs within the delimited area of the Alps. The testing area of Switzerland is marked as well. Green, orange and pink colors distinguish the proposed training, validation and testing sets, respectively. The query photographs are paired with our synthetically rendered databases. The *sparse database* contains rendered views at positions identical to the query positions. The *uniform database* has positions defined by a uniform grid with 500 m spatial sampling across the whole Alps area. **Zoom 1** provides a closer look. For simplicity of illustration, only the testing area is densely covered by the uniform grid of database positions (purple). To speed up the training process, we additionally provide the *uniform compact database* used (only) for training (dark green), where only grid positions that are within 1 kilometer from any query position are kept. **Zoom 2** offers a detailed look at a specific position, where a query photograph is complemented by rendered image modalities (semantic segmentations, silhouette maps and depth maps). There is a total of 12 views (3 shown) at each position for each modality, covering the 360° field of view.

3.3. Uniform dataset

The *uniform database* is designed to enable the real-world task of localization across a large natural area – the Alps. The geographical positions included inside the *uniform database* are defined by a dense uniform grid with a step of 500 meters at both axes. As the database covers approximately 250 000 square kilometers, this results in nearly 1 million positions. As in the case of the *sparse database*, at each of these positions we render 12 synthetic views of 500×500 resolution, each with 60° field of view and 30° separation in yaw angle. This results in nearly 12 million database views inside the *uniform database*.

When using the *uniform database*, the localization task is much more challenging than in the case of the *sparse database*. Not only is the number of potential database candidates much bigger, but the database positions are not adjusted to any distribution of query positions. Therefore, localization methods have to exhibit a sufficient level of robustness. It also means that this database can be paired with any dataset of query photographs from the Alps region.

We propose to pair the *uniform database* with the "CrossLocate" query dataset of 12353 query photographs described in Sec. 3.1, forming the *Uniform dataset*. Consistently with the *Sparse dataset*, we propose a split into training, validation and testing sets, with the area of Switzerland put aside for the testing set, as summarized in Tab. 2.

Working with all 8 million database images assigned to the training set can lead to excessive processing times. To make this process faster, we propose a *compact* version of the training part of the *uniform database*. In the *compact uniform database* (Fig. 2, zoom level 1), the training database positions are filtered so that only the grid positions that are within 1 kilometer from some query position are kept. This drastically reduces the size of the training database, while preserving all the database images that might be required to construct ("positive") pairs of queries and their corresponding database counterparts.

Uniform dataset	train	val	test	total
uniform database	7.85M	0.89M	1.98M	**10.72M**
uni. comp. database	**161K**	0.89M	1.98M	3.03M
queries (CrossLocate)	8324	516	3513	**12353**

Table 2. Numbers of query and database images in the *Uniform dataset*, and their split into training, validation and testing sets.

4. CrossLocate method

To cope with the low amount of available data, *i.e.* user-taken photographs, and its sparse and uneven distribution across natural environments, we propose a *retrieval-like* localization method that builds a powerful image representa-

tion. We design our localization method as a *cross-modal image retrieval* and aim to represent each image (place) by a single *global* descriptor. This representation is learned automatically in an end-to-end manner. Each component of our architecture is carefully selected and empirically validated (see Sec. 5.4).

4.1. Architecture

The basis for the architecture of our method are standard convolutional blocks. We use 5 convolutional blocks, with each block consisting of 2-3 convolutional layers with ReLU units, and each block is ended with max-pooling. In the last block, we do not use any pooling, and we also do not use the ReLU activation in the very last convolutional layer in order to not restrict the resulting representation to be non-negative. A detailed description of our architecture is available in the supplementary material. Assuming a (three-channel) input image I, we obtain a 3D activation tensor $T \in R^{H \times W \times D}$ seen as $H \times W$ D-dimensional features. We apply channel-wise L2 normalization to this tensor at each of the $H \times W$ spatial positions separately.

To produce a single global descriptor as a representation for each image, we end the architecture with a global max-pooling layer, and apply another L2 normalization. The resulting descriptor has $D = 512$ dimensions. As illustrated in Sec. 5.4, the replacement of any of these components leads to a significant drop in localization performance. We specifically stress the importance of the final max-pooling in our cross-modal task.

Our single (single-branch) model is capable of extracting the deep representations for both query RGB photographs and rendered database views. When working with multiple database modalities, each modality takes one input channel.

4.2. Training process

We initialize our architecture with weights pretrained on the ImageNet dataset [16] to better cope with the low number of available images and to combat overfitting. We work with input images scaled to a unified resolution of 500×500 pixels. This resolution corresponds to the 60° field of view covered by the database views. Therefore, we scale the content of each query photograph according to its actual field of view. For example, the useful content of a query photograph with 30° field of view would be 225×225 pixels. In this way, we preserve the correct scale of the scene and enable precise localization. Taking different scales into consideration is one of the key aspects that differentiates localization from general retrieval.

We train our method using a variant of the triplet loss objective [36], similar to [3], *i.e.* training is done by presenting the method with triplets of so-called query (anchor) images together with corresponding positive and negative examples. This loss function is combined with the euclidean metric, which measures the distance between extracted image representations. The goal is to learn a representation where the distances between a query and its positive example(s) are smaller than the distances between the query and its negative examples.

We provide technical details and thoroughly describe our three-step supervision process of selecting the positive and negative examples in the supplementary material (Sec. 1).

5. Experiments

5.1. Evaluation protocol

We measure the performance of all considered methods through the *recall metric*. For each query, a specific number N of the nearest database candidates is retrieved, and we measure the percentage of successfully localized query images (recall). A query is considered successfully localized if at least one of the retrieved database candidates is within a specific location distance from the query ground-truth. We provide results for a wide range of location thresholds, but emphasize the results at 1 kilometer tolerance.

Since existing methods typically evaluate only the correctness of the position estimate and ignore the correctness of the estimated orientation, we do not consider the orientation (yaw angle). However, we show that incorporating an orientation restriction (30°) results only in a negligible performance decrease for our approach. Our main goal is to measure the capabilities of our method as a standalone approach. Therefore, we mainly report the recall at $N = 1$ database candidate retrieved for each query image. However, retrieval methods may also be followed by a geometric verification (reranking) of the retrieved candidates. Accordingly, we report the recall at $N = 100$ database candidates retrieved for each query too.

5.2. Testing sets

We evaluate our and existing methods mainly on the testing sets of the *Sparse dataset* (Sec. 3.2) and *Uniform dataset* (Sec. 3.3). The *uniform database* can also be combined with other datasets of query photographs. Specifically, we use the CH1 dataset [35] of 203 photographs. We also use the CH2 dataset [35] of 949 photographs with known positions. The proposed pairing between queries and databases for the purpose of testing is shown in Tab. 3.

Testing sets	query	database
Sparse dataset	1186	14232 (*sparse database*)
Uniform dataset	3513	1979160 (*uniform database*)
CH1 [35]	203	1979160 (*uniform database*)
CH2 [35]	949	1979160 (*uniform database*)

Table 3. Summary of testing sets used for evaluation and comparisons with other methods.

5.3. Modalities evaluation

To provide an insight into the localization potential of image modalities, we measure the localization performance with the individual database modalities. We also measure results for horizon lines (used in [35]), which we derive from semantic segmentations. A result for the combination of the three base modalities is also included. All the experiments were trained on the *Sparse dataset* (Sec. 3.2), using the same database modality both in training and testing.

Image modalities. Fig. 3 shows the results for the evaluation on the testing sets of the *Uniform dataset* (left) and *Sparse dataset* (right). The worst performance is generally obtained with horizon lines and semantic segmentations. This aligns with our expectations, as horizon lines intuitively contain the least amount of information. While semantic segmentations theoretically contain more information in the form of spatial segments, we attribute their weak performance to the bad correspondence between the ideal synthetic segments and the real contents of query photographs, whose appearance is very variable. Silhouette maps provide slightly better results as they carry information about all terrain discontinuities, not limited to horizon lines. The best performance by a significant margin is achieved with depth maps. This shall be intuitive, as they not only include all the information from horizon lines and silhouette maps, but also offer depth information in each pixel. The combination of segments, silhouettes and depth maps (`seg-sil-dep`) achieves only a minor improvement on the *Sparse dataset*, which is in agreement with the mentioned reasoning. The only exception to the described behavior is observable on the CH1 dataset, where horizon lines are nearly as good as depth maps (shown in the supplementary material). Therefore, we use **depths maps as the database image modality** in the rest of the experiments.

Synthetic modalities. We performed identical experiments with the individual modalities in a fully synthetic scenario, where the query photographs were replaced with synthetic views at the corresponding locations (not shown in plots). These synthetic query images are part of the GeoPose3K dataset [11]. While the depth maps prove to be the best choice again, the difference between the individual modalities is much less apparent and often negligible. This suggests that the gap between single-domain problems (*e.g.* synthetic data only) and cross-domain problems is significant, as the observations cannot be directly transferred between these tasks.

Synthetic pre-training. In order to further support the previous statement, we used the models trained in the fully synthetic scenario as an initialization point for our training with real photographs (used as queries). This does not bring any performance improvement compared to the initialization based on the ImageNet dataset. This suggests that the synthetic pre-training does not offer additional knowledge that could be utilized in our cross-modal task.

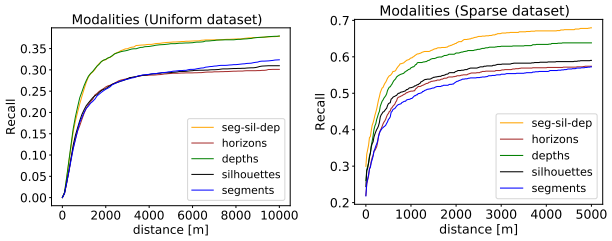

Figure 3. Localization performance of different image modalities. Results are measured on the testing sets of the *Uniform dataset* (left) and *Sparse dataset* (right).

Cross-modal training. The insufficiency of any of the single-domain initializations, together with the importance of our cross-modal training, is further apparent from Fig. 5. We measure the localization performance when models are only initialized and directly evaluated, without our cross-modal training. Both the intialization based on the ImageNet dataset (`ImageNet`) and the initialization based on pretraining on the synthetic datasets (not shown) exhibit weak localization performance compared to the results obtained after our cross-modal training (`Our`).

5.4. Architecture components evaluation

To emphasize the importance of the design decisions related to our method's architecture, we provide results obtained when replacing our base convolutional architecture with a different one, or when removing the important components of our solution. All the experiments were trained on the *Sparse dataset* and use photographs as the queries and depth maps as the database image modality. Results for the testing set of the *Sparse dataset* are shown in Fig. 4.

Base architecture. We replaced our base architecture with the simple architecture of AlexNet [28], as well as multiple variants of ResNets [23] to represent more advanced architectures. None of these provided better results. Apart from the difference in the architecture sizes, the worse results could be attributed to the local response normalizations of AlexNet and the batch normalizations of ResNets. These normalizations are not suitable for the cross-modal character of our data (photographs and rendered views).

Two-branch architecture. Surprisingly, we found only a minor benefit in using a two-branch modification of our architecture (`Two-branch`). This is an interesting result as cross-view and cross-modal approaches tend to employ two-branch architectures [25, 30, 13]. Unfortunately, they typically do not provide results for a single-branch variant. According to our observations, the performance of the cross-modal approach is much more dependent on the choice of an appropriate architecture (both in terms of layer types and numbers of parameters) than on providing a separate branch for each input modality.

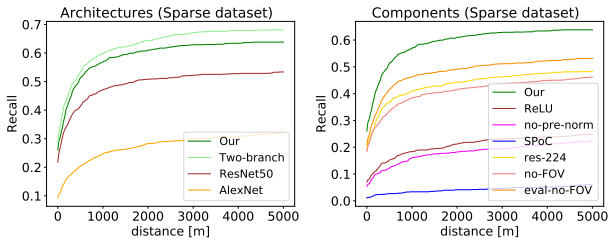

Figure 4. Importance of individual design choices in our approach. Results measured on the testing set of the *Sparse dataset*.

Deep representation. Both the first L2 normalization and the subsequent global max-pooling proved to be crucial to enable our cross-modal localization. The often-used sum-pooling [45] does not work in our cross-modal scenario. In addition, the removal of the ReLU activation from the last convolutional layer is important for the richness of the resulting deep representation. Individual results obtained while keeping the final ReLU unit (`ReLU`), removing the first L2 normalization (`no-pre-norm`) or using sum-pooling (`SPoC`), are shown in Fig. 4. We also experimented with the NetVLAD aggregation, however, the required dimensionality reduction, combined with the low amount of training data, led to a generally bad performance in our task.

Input resolution. Our choice of higher input image resolution also leads to better results than those obtained with the usual input resolution of 224×224 pixels (`res-224`). This is mainly due to the input images being scaled according to their field of view, which might result in the downscaling of the image content. Our increased resolution helps balance out this effect.

Field of view information. We also demonstrate the benefit of scaling the content of input images according to their field of view, in comparison to discarding the field of view information and scaling all the images so that they fit the input resolution. Fig. 4 shows the clear benefit of working with scaled images, as discarding this information in training and evaluation leads to a drop in performance (`no-FOV`). However, while the scaling is beneficial for training, it is not essential for evaluation, as our method still exhibits satisfactory performance when the field of view is not known (`eval-no-FOV`).

6. Comparison with state-of-the-art methods

We compare our approach with other methods on the task of localization of photographs captured within the mountainous areas across Switzerland. We provide results for the HLoc method [35], which uses horizon lines for localization. We further show results for state-of-the-art single-domain retrieval methods DELG [14] and HOW [39]. The methods were trained on the *Uniform dataset* with 8324 query photographs and depth maps were used as the database image modality.

6.1. Setup of competing methods

HLoc [35] localizes photographs by first detecting and encoding horizon lines. Subsequently, it matches them with a uniform database of synthetic horizon lines, which are extracted from a digital elevation model. As the authors [35] did not provide the code, we used the published re-implementation [11]. For a fair comparison, the geographic positions used for the creation of the synthetic horizon lines are the same as the positions in the *Uniform dataset*. The original work relied on the user support and guidance for successful detection of (query) horizon lines. The authors only provided such semi-automatically extracted horizon lines for the CH1 dataset. To examine the method's behavior on other datasets, and in a fully automatic scenario, we used a state-of-the-art segmentation method Deeplab v3+ [15] and a dedicated Edge-Less detection algorithm [2] to detect horizon lines for the query photographs of the *Uniform dataset*, as well as the CH1 and CH2 datasets. We show (see the supplementary material) that this automatic detection does not lead to a dramatic decrease in the localization performance compared to the human guided detection on CH1.

Although the *DELG method* [14] is defined as a retrieval, it is trained as a classification task. Therefore, we needed to adapt our *Uniform dataset* in the corresponding way. We created geographical clusters containing both query photographs and rendered database views from the same locations while also sharing a similar orientation. The clusters have a radius of 1 kilometer and views are within 15° from one another. Each cluster is represented by one class for a total of 2598 classes used during the training.

The *HOW method* [39] was originally trained in a similar way to our approach, as it utilized positive and negative examples for each query image in combination with a contrastive loss. However, it is designed to work with datasets created by structure-from-motion, where 3D point correspondences between images are known and used for the selection of positive and negative examples. To enable the comparison, we replaced their supervision with our localization supervision which provides training examples including augmentation. This puts both methods on a common ground, leaving emphasis on the method architectures.

6.2. Discussion

We show detailed results for the comparison with our method on the testing set of the *Uniform dataset* in Fig. 5. Additional results on the *Sparse*, CH1 and CH2 datasets are available in the supplementary material (Sec. 3).

For the evaluation, we provide recall values at 1 database candidate retrieved for each query (recall@1, plotted as solid lines), and also recall values at 100 database candidates retrieved (recall@100, plotted as dashed lines). Recall@100 represents the best possible results obtainable

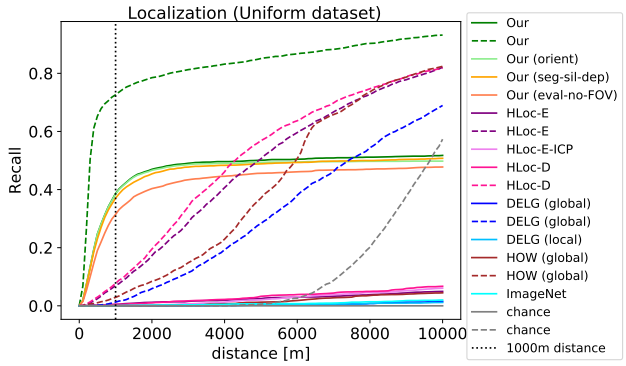

Figure 5. Localization performance of our approach and state-of-the-art methods. Solid lines: recall at 1 database candidates retrieved for each query, dashed lines: recall at 100 candidates. Our approach outperforms the state-of-the-art methods in both cases.

with a perfect geometric verification applied to the top 100 database candidates of each query.

For DELG, we provide results for the retrieval step based on global descriptors (`DELG (global)`) and for the geometric verification with local descriptors (`DELG (local)`). For HOW, we show retrieval results with global descriptors (`HOW (global)`). We provide results for ASMK aggregation used in HOW only on the *Sparse dataset* (`HOW (ASMK)`), because of the high memory requirements of ASMK descriptors in combination with our *Uniform dataset*. For HLoc, we provide results for the retrieval step done based on (query) horizon lines detected by the Edgeless method (`HLoc-E`) and Deeplab method (`HLoc-D`).

The results for recall@1 (solid lines) show that our CrossLocate approach (`Our`) is the only one among the competing methods reaching a reasonable localization performance on the challenging *Uniform dataset*.

In our setup, the accuracy reached by DELG on the validation classification set did not translate to the performance during the retrieval evaluation on the testing set. We attribute this mainly to the unsuitability of the classification training for localization in natural environments, where data is scarce. Classes created from such data contain only low numbers of images which make it difficult for the method to properly learn the appearance of a given place, and to build a robust representation as a result.

We attribute the failure of the HOW approach to the different architecture decisions. We show that our L2 normalization of activations, omitted ReLU unit, as well as max-pooling are crucial for our cross-modal localization, while the sum-pooling (used in HOW) fails. In addition, the choice of the ResNet backbone architecture with batch normalizations in HOW is not appropriate for our cross-modal task. Both HOW and DELG were originally trained on hundreds of thousands of images. Our method does not require such a large dataset.

Among the compared state-of-the-art methods, the HLoc approach performs the best. However, its results are still weak, hinting at the low amount of information available in horizon lines when localizing the challenging query photographs of the *Uniform dataset* across hundreds of thousands of square kilometers and millions of places.

For our CrossLocate approach, the recall@1 significantly outperforms the recall@100 of other methods, up to approximately a 5 kilometer localization tolerance. Beyond this threshold, the other methods arguably benefit from randomly guessing correct places. At the strict 1 kilometer threshold, we report 38.66% recall@1 and 72.62% recall@100. Additional values, as well as qualitative evaluation with localization examples, can be found in the supplementary material (Sec. 4). We also provide a result for the combination of the database modalities (`Our (seg-sil-dep)`), which brings no benefit compared to our choice of depth maps. Furthermore, we evaluate our approach when taking orientation into consideration (`Our (orient)`). When correct localization is required to be within 30° from ground-truth yaw angle, the performance of our method is nearly unchanged. We also show that discarding the field of view information used for the scaling of query photographs leads only to a small decrease in performance on the challenging *Uniform dataset* (`Our (eval-no-FOV)`). Finally, we provide results for the case when our method is initialized based only on the ImageNet dataset (`ImageNet`), but our cross-modal training is not performed. Results for obtaining database candidates for each query by random guessing (`chance`) are also shown.

7. Conclusions

We addressed multiple aspects of the open and challenging task of visual geo-localization in natural environments. We built a large benchmark dataset of various rendered image modalities spanning the whole range of Alps, which opens the way for future research in the field. The dataset enabled our ablation studies, where we provided an insight into the difficulty of this cross-modal task, as well as into the important specifics of our architecture. The analysis of the modalities revealed a profound effect of depth information on localization performance. Finally, we introduced a cross-modal retrieval-based localization method, which matches query photographs to rendered databases and delivers state-of-the-art results.

Acknowledgements. This work was supported by project *LTAIZ19004 Deep-Learning Approach to Topographical Image Analysis*; by the Ministry of Education, Youth and Sports of the Czech Republic within the activity INTER-EXCELENCE (LT), subactivity INTER-ACTION (LTA), ID: SMSM2019LTAIZ. Computational resources were partly supplied by the project *"e-Infrastruktura CZ" (e-INFRA CZ ID:90140)* supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

[1] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz, and Richard Szeliski. Building Rome in a day. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 72–79, 2009.

[2] Touqeer Ahmad, George Bebis, Monica N. Nicolescu, Ara V. Nefian, and Terry Fong. An edge-less approach to horizon line detection. In *14th IEEE International Conference on Machine Learning and Applications, ICMLA 2015, Miami, FL, USA, December 9-11, 2015*, pages 1095–1102. IEEE, 2015.

[3] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307. Washington, D.C., USA: IEEE Computer Society Press, 2016.

[4] Anil Armagan, Martin Hirzer, Peter M. Roth, and Vincent Lepetit. Learning to align semantic segmentation and 2.5D maps for geolocalization. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4590–4597. IEEE, 2017.

[5] Roberto Arroyo, Pablo F. Alcantarilla, Luis M. Bergasa, and Eduardo Romera. Fusion and binarization of CNN features for robust topological localization across seasons. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4656–4663. IEEE, 2016.

[6] Georges Baatz, Olivier Saurer, Kevin Köser, and Marc Pollefeys. Large scale visual geo-localization of images in mountainous terrain. In *Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II*, pages 517–530. Springer-Verlag Berlin Heidelberg, 2012.

[7] Lionel Baboud, Martin Čadík, Elmar Eisemann, and Hans Peter Seidel. Automatic photo-to-terrain alignment for the annotation of mountain pictures. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–48. IEEE Computer Society, 2011.

[8] Assia Benbihi, Stephanie Arravechia, Matthieu Geist, and Cédric Pradalier. Image-based place recognition on bucolic environment across seasons from semantic edge description. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3032–3038, 2020.

[9] Eric Brachmann and Carsten Rother. Learning Less is More - 6D Camera Localization via 3D Surface Regression. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4654–4662. IEEE, 2018.

[10] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-Aware Learning of Maps for Camera Localization. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2616–2625. IEEE, 2018.

[11] Jan Brejcha and Martin Čadík. GeoPose3K: Mountain landscape dataset for camera pose estimation in outdoor environments. *Image and Vision Computing*, 66:1–14, 2017.

[12] Jan Brejcha and Martin Čadík. Camera orientation estimation in natural scenes using semantic cues. In *Proceedings - 2018 International Conference on 3D Vision*, pages 208–217. IEEE, 2018.

[13] Jan Brejcha, Michal Lukáč, Yannick Hold-Geoffroy, Oliver Wang, and Martin Čadík. LandscapeAR: Large scale outdoor augmented reality by matching photographs with terrain models using learned descriptors. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX*, pages 295–312, Cham, 2020. Springer International Publishing.

[14] Bingyi Cao, André Araujo, and Jack Sim. Unifying deep local and global features for image search. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*, pages 726–743, Cham, 2020. Springer International Publishing.

[15] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[17] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In *Computer Vision – ECCV 2020*, volume 12349 of *Lecture Notes in Computer Science*, pages 369–386, Cham, Germany, 2020. Springer International Publishing.

[18] Marcel Geppert, Peidong Liu, Zhaopeng Cui, Marc Pollefeys, and Torsten Sattler. Efficient 2d-3d matching for multi-camera visual localization. In *ICRA*, pages 5972–5978. IEEE, 2019.

[19] Petr Gronát, Josef Sivic, Guillaume Obozinski, and Tomas Pajdla. Learning and Calibrating Per-Location Classifiers for Visual Place Recognition. *International Journal of Computer Vision*, 118(3):319–336, 2016.

[20] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003.

[21] James Hays and Alexei A Efros. IM2GPS: Estimating geographic information from a single image. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. New York, NY, USA: IEEE, 2008.

[22] James Hays and Alexei A. Efros. Large-scale image geolocalization. In *Multimodal Location Estimation of Videos and Images*, chapter Large-scale image geolocalization, pages 41–62. Springer International Publishing, 2015.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[24] Jared Heinly, Johannes L Sch, Enrique Dunn, and Jan-michael Frahm. Reconstructing the World in Six Days. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[25] Sixing Hu, Mengdan Feng, Rang M.H. Nguyen, and Gim Hee Lee. CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7258–7267. IEEE, 2018.

[26] Arnold Irschara, Christopher Zach, Jan Michael Frahm, and Horst Bischof. From structure-from-motion point clouds to fast location recognition. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, pages 2599–2606, 2009.

[27] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*, pages 2938–2946. New York, NY, USA: IEEE, 2015.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25(2), 2012.

[29] Tsung Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898. IEEE, 2013.

[30] Tsung Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5007–5015. IEEE, 2015.

[31] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[32] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-Scale Image Retrieval with Attentive Deep Local Features. In *Proceedings of the 2017 IEEE International Conference on Computer Vision*, pages 3476–3485. IEEE, 2017.

[33] Lorenzo Porzi, Samuel Rota Bulò, Oswald Lanz, Paolo Valigi, and Elisa Ricci. An automatic image-to-DEM alignment approach for annotating mountains pictures on a smartphone. *Machine Vision and Applications*, 28(1-2/2017):101–115, 2016.

[34] Filip Radenović, Giorgos Tolias, and Ondřej Chum. CNN image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I*, pages 3–20, 2016.

[35] Olivier Saurer, Georges Baatz, Kevin Köser, L'ubor Ladický, and Marc Pollefeys. Image Based Geo-localization in the Alps. *International Journal of Computer Vision*, 116(3):213–225, 2016.

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[37] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. In *Advances in Neural Information Processing Systems 32*, pages 10090–10100. Curran Associates, Inc., 2019.

[38] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. To aggregate or not to aggregate: Selective match kernels for image search. In *2013 IEEE International Conference on Computer Vision*, 2013.

[39] Giorgos Tolias, Tomas Jenicek, and Ondřej Chum. Learning and aggregating deep local descriptors for instance-level recognition. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, 2020.

[40] Akihiko Torii, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. Visual Place Recognition with Repetitive Structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2346–2359, 2015.

[41] Nam Vo, Nathan Jacobs, and James Hays. Revisiting IM2GPS in the Deep Learning Era. In *Proceedings of the 2017 IEEE International Conference on Computer Vision*, pages 2640–2649. IEEE, 2017.

[42] Nam N. Vo and James Hays. Localizing and orienting street views using overhead imagery. In *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I*. Springer International Publishing, 2016.

[43] Peng Wang, Ruigang Yang, Binbin Cao, Wei Xu, and Yuanqing Lin. DeLS-3D: Deep Localization and Segmentation with a 3D Semantic Map. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5860–5869. IEEE, 2018.

[44] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet - photo geolocation with convolutional neural networks. In *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 37–55. Cham: Springer International Publishing, 2016.

[45] Artem Babenko Yandex and Victor Lempitsky. Aggregating Deep Convolutional Features for Image Retrieval. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*, pages 1269–1277. IEEE, 2015.

[46] Bernhard Zeisl, Torsten Sattler, and Marc Pollefeys. Camera Pose Voting for Large-Scale Image-Based Localization. In *2015 IEEE International Conference on Computer Vision*, pages 2704–2712, 2015.