Visual Localization using Imperfect 3D Models from the Internet

Vojtech Panek\textsuperscript{1,2} Zuzana Kukelova\textsuperscript{3} Torsten Sattler\textsuperscript{2}

\textsuperscript{1} Faculty of Electrical Engineering, Czech Technical University in Prague
\textsuperscript{2} Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague
\textsuperscript{3} Visual Recognition Group, Faculty of Electrical Engineering, Czech Technical University in Prague

\{vojtech.panek,torsten.sattler\}@cvut.cz kukelzuz@fel.cvut.cz

Abstract

Visual localization is a core component in many applications, including augmented reality (AR). Localization algorithms compute the camera pose of a query image w.r.t. a scene representation, which is typically built from images. This often requires capturing and storing large amounts of data, followed by running Structure-from-Motion (SfM) algorithms. An interesting, and underexplored, source of data for building scene representations are 3D models that are readily available on the Internet, e.g., hand-drawn CAD models, 3D models generated from building footprints, or from aerial images. These models allow to perform visual localization right away without the time-consuming scene capturing and model building steps. Yet, it also comes with challenges as the available 3D models are often imperfect reflections of reality. E.g., the models might only have generic or no textures at all, might only provide a simple approximation of the scene geometry, or might be stretched. This paper studies how the imperfections of these models affect localization accuracy. We create a new benchmark for this task and provide a detailed experimental evaluation based on multiple 3D models per scene. We show that 3D models from the Internet show promise as an easy-to-obtain scene capturing and model building steps. Yet, it also comes with challenges as the available 3D models are often imperfect reflections of reality. E.g., the models might only have generic or no textures at all, might only provide a simple approximation of the scene geometry, or might be stretched.

1. Introduction

Visual localization is the task of estimating the precise position and orientation, i.e., the camera pose, from which a given query image was taken. Localization is a core capability for many applications, including self-driving cars\cite{35}, drones\cite{54}, and augmented reality (AR)\cite{16,57}.

Visual localization algorithms mainly differ in the way they represent the scene, e.g., explicitly as a 3D model\cite{27,28,42,53,56,57,73,75,76,78,83,92,107} or a set of posed images\cite{10,66,108,110}, or implicitly via the weights of a neural network\cite{5,12–14,17,46,47,59,101,104}, and in the way they estimate the camera pose, e.g., based on 2D-3D\cite{27,28,73,75,78,92,107} or 2D-2D\cite{10,110} matches or as a weighted combination of a set of base poses\cite{46,47,59,66,81,104}. Existing visual localization approaches typically use RGB(-D) images to construct their scene representations. Yet, capturing sufficiently many images and estimating their camera poses, e.g., via Structure-from-Motion (SfM)\cite{82,90}, are challenging tasks by themselves, especially for non-technical users unfamiliar with how to best capture images for the reconstruction task.

As shown in\cite{4,63,85,88,109}, modern local features such as\cite{21,23,91,111} are capable of matching real images with renderings of 3D models, even though they were never explicitly trained for this task. This opens up a new way of obtaining the scene representations needed for visual localization: rather than building the representation from images captured in the scene, e.g., obtained via a service such as Google Street View, crowd-sourcing\cite{99}, or from photo-sharing websites\cite{52}, one can simply download a ready-made 3D model from the Internet (cf. Fig. 1), e.g., from 3D model sharing websites such as Sketchfab and 3D Warehouse. This removes the need to run a SfM system such as COLMAP\cite{82} or RealityCapture on image data, which can be a very complicated step, especially for non-experts such
as artists designing AR applications. As such, this approach has the potential to significantly simplify the deployment of visual localization-based applications.

However, using readily available 3D models from the Internet to define the scene representation also comes with its own set of challenges (cf. Fig. 2): (1) **fidelity of appearance**: the 3D models might not be colored / textured, thus resulting in very abstract representations that are hard to match to real images [63]. Even if a model is textured, the texture might be generic and repetitive rather than based on the real appearance of the scene. In other cases, the texture might be based on real images, but severely distorted or stretched if these images were captured from drones or planes. (2) **fidelity of geometry**: some 3D models might be obtained via SfM and Multi-View Stereo (MVS), resulting in 3D models that accurately represent the underlying scene geometry. Yet, this does not always need to be the case. For example some models might be obtained by extruding building outlines, resulting in a very coarse model of the scene geometry. Others might be created by artists by hand, resulting in visually plausible models with overly simplified geometry, or with wrong aspect ratios, e.g., a model might be too high compared to the width of the building.

Naturally, the imperfections listed above negatively affect localization accuracy. The goal of this work is to quantify the relation between model inaccuracy and localization accuracy. This will inform AR application designers which 3D models are likely to provide precise pose estimates. Looking at Fig. 2, humans seem to be able to establish correspondences between the models, even if the models are very coarse and untextured. Similarly, humans are able to point out correspondences between coarse and untextured models and real images that can be used for pose estimation in the context of visual localization. As such, we expect that it is possible to teach the same to a machine. We thus hope that this paper will help researchers to develop algorithms that close the gap between human and machine performance for this challenging matching task.

In summary, this paper makes the following contributions: (1) we introduce the challenging and interesting tasks of visual localization w.r.t. to 3D models downloaded from the Internet. (2) we provide a new benchmark for this task that includes multiple scenes and 3D models of different levels of fidelity of appearance and geometry. (3) we present detailed experiments evaluating how these different levels of fidelity affect localization performance. We show that 3D models from the Internet represent a promising new category of scene representations. Still, our results show that there is significant room for improvement, especially for less realistic models (which often are very compact to store). (4) We make our benchmark publicly available (v-pnk.github.io/cadloc) to foster research on visual localization algorithms capable of handling this challenging task.

2. Related Work

**Localization based on images.** Currently, state-of-the-art localization approaches that can handle strong changes in scene appearance, e.g., day-night and seasonal changes [79, 96], are based on local features [15,41,42,53,58,64,73,75,77,78,83,92–94,107]. They typically represent the scene through Structure-from-Motion (SfM) point clouds, where each 3D point is associated with 2D image features from the database images used to reconstruct the point cloud. 2D-3D matches between features extracted in the query images and 3D points in the scene model are then used to estimate the query pose by applying a minimal solver [49,65] inside a RANSAC scheme [6–9,20,24,51,69]. Hierarchical localization approaches [41,42,74,80,93,94] use an intermediate image retrieval [1,29,44,71,89,98] step to focus on smaller parts of the scene for 2D-3D matching, thus increasing the scalability of localization techniques. Rather than using sparse SfM point clouds, dense meshes can also be used as a scene representation [63]. The scene geometry can also be represented implicitly through the weights of machine learning models [12–14,17–19,87,100].

An alternative to using a 3D model is to only store a database of images with known camera poses. The camera pose of the query can then be either approximated based on the poses of database images retrieved via image retrieval [1,29,44,71,89,98], or accurately estimated from 2D-2D matches between the query and the retrieved images [10,110]. It is further possible to implicitly store the database images by training neural networks that regress the camera pose of a query image w.r.t. the database images [5,22,46,47,50,59,61,62,86,104].

Common to these approaches is that their scene representation is derived from images and thus accurately models the scene geometry and appearance. We investigate how inaccurate approximations of the scene geometry and appearance, in the form of 3D models downloaded from the Internet, impact feature-based localization performance.

**Localization using multi-modal data.** While less popular than sparse SfM-based representations, dense meshes have been used in the visual localization literature [3,4,15,31,60,63,67,70,85,88,97,102,109]. Prior work can roughly be divided into methods using specialized strategies to align real images to (potentially coarse) 3D models [31,67,70], e.g., using contours [67] or skylines [70], and works based on matching local features [3,4,15,60,63,85,88,97,109]. The later line of work has shown that modern local features can match real photos against non-photorealistic renderings of colored meshes or even against meshes without any color [15,63,97]. This observation has motivated our work on using 3D meshes downloaded from the Internet, with potentially non-realistic appearance, as scene representations for visual localization. Prior work in this direc-
tion typically used as geometrically-precise 3D models as possible. In contrast, we investigate to what degree distortions in model geometry and appearance can be handled as we have no control over the quality of 3D models available on the Internet. To the best of our knowledge, we are the first who, in detail, study how the imperfections of Internet models affect localization accuracy.

CAD and other 3D models are also commonly used by object pose estimation approaches [2, 26, 30, 32, 36–38, 48, 68]. These methods learn to establish correspondences between real images and precise models of object instances or 3D models depicting the same class of objects. Such instance- or class-specific training is not applicable in our scenario due to a lack of scene-specific training data.

Cross-view matching estimates the camera pose based on correspondences found between the ground-level query image and a 2D bird’s eye view map (such as a satellite image or a semantic landscape map) [39, 40, 55, 85, 103, 106]. Similar to our problem setting, they also need to solve a challenging matching task between different modalities. As in our approach, these maps can often be obtained from the Internet without visiting the place. On the other hand, the maps capture the actual state of the scene, although the low resolution of such maps can be a similar challenge as the low geometric and appearance fidelity of our 3D models.

3. Datasets

The goal of this work is to study how and to what degree inaccurate 3D models downloaded from the Internet affect visual localization performance. We are motivated by the observation that modern local features, such as the ones used in state-of-the-art localization pipelines [63, 73, 75, 91], can establish correspondences between a real image and renderings of 3D models [15, 63, 109]. This allows us to apply a state-of-the-art localization pipeline [63] on renderings of different 3D models. While prior work focused on 3D models obtained from images [63, 109], i.e., models that
For our experimental evaluation, we collected 3D models for 6 scenes (cf. Fig. 2 and Tab. 1). The scenes were selected based on the availability of interesting 3D models covering multiple challenges, e.g., different levels of appearance and geometric fidelity, and the availability of query images. For the 3D models, we distinguish between models that were automatically obtained from images via MVS and CAD models that were created manually. The following describes the models used per scene. We publish our benchmark (v-pnk.github.io/cadloc) to foster research on this interesting and challenging topic. In particular, our benchmark can be used to measure how well local features are able to handle complex matching tasks between real images and more abstract representations of the scene.

For each scene, we collected a set of query images with known poses (in a coordinate system with an arbitrary scale). We then aligned the models downloaded from the Internet to the coordinate system of the images. This allows us to measure the accuracy of the camera poses estimated by the localization algorithm in a common frame and so we decided to use just their geometry (cf. supp.).

accurately align with reality, this work aims to answer the question to which degree deviations from reality, both in terms of geometry and appearance, can be handled.

For our experimental evaluation, we collected 3D models for 6 scenes (cf. Fig. 2 and Tab. 1). The scenes were selected based on the availability of interesting 3D models covering multiple challenges, e.g., different levels of appearance and geometric fidelity, and the availability of query images. For the 3D models, we distinguish between models that were automatically obtained from images via MVS and CAD models that were created manually. The following describes the models used per scene. We publish our benchmark (v-pnk.github.io/cadloc) to foster research on this interesting and challenging topic. In particular, our benchmark can be used to measure how well local features are able to handle complex matching tasks between real images and more abstract representations of the scene.

For each scene, we collected a set of query images with known poses (in a coordinate system with an arbitrary scale). We then aligned the models downloaded from the Internet to the coordinate system of the images. This allows us to measure the accuracy of the camera poses estimated by the localization algorithm in a common frame of reference and thus to compare pose accuracy between the models. Unless mentioned otherwise, we used images from the Image Matching Challenge (IMC) 2021 as queries. The images are part of the YFCC100M dataset [95], depicting popular landmarks at different points in time. The IMC provides the camera poses from [34] for each image. Using these poses, we created a MVS mesh using RealityCapture and aligned the Internet models with this mesh, using ICP-based refinement [72] starting from a manual initialization.

**Notre Dame (Front Facade).** We selected 7 models of Notre Dame from the Internet, representing different levels of geometric detail and fidelity of appearance. Model A is the only one coming from photogrammetry reconstruction, the others are CAD models created manually. The models B, C and D are textured. While the texture of the model B looks rather realistic, the texture of C seems to come from a painting and the texture of D is a combination of low-quality photos and generic textures. The geometric detail is the highest for the model B and the lowest in the case of the model D, where the front facade of the building is essentially only planar. Model C is somewhere in between. E has a very high level of geometric detail, F is comparable to the level of detail of B and the geometry of D and G is identical. E, F and G do not contain any color information.

**Pantheon (Exterior).** Here we use five models: Model A was created by photogrammetry. Model B contains realistically looking textures. Models C and D originally contained generic textures, which did not correspond to reality in any way and so we decided to use just their geometry (cf. supp. mat.). C contains a very high level of geometric detail and adds on top multiple details, e.g., statues, which the real building does not contain in its current state. Model D has a medium level of geometric detail. Model E is constructed from a set of voxels and contains non-realistic coloring.

**Reichstag.** We have four textured CAD models of Reichstag (A, B, C, D). Models A and B depict the current state of the building, while C and D are trying to create a historic depiction. Models E and F do not use texture. Models A, C and E have a higher level of geometric detail than the others.

**Table 1.** List of scenes and 3D models used for the evaluation. The query images for the scenes were obtained from the Image Matching Challenge (IMC) 2021 [34, 45, 95], the Aachen Day-Night v1.1 dataset [79, 80, 109], and our own recordings. We distinguish between models directly created from images via MVS and models created from human input (CAD).
**St. Peter’s Square.** We use four models: Model A is generated by photogrammetry and also contains surrounding buildings. Models B and C have realistic textures. All models have a comparably high level of geometric detail.

**St. Vitus Cathedral.** We use four models: an MVS model (B) containing a larger area, reconstructed from drone footage via photogrammetry, a version of this model that only contains the cathedral (A), and two CAD models (C and D) created using SketchUp. Models C and D were chosen as they provide different levels of detail, with model C containing finer geometric details that are missing in model D. Both C and D use low-resolution textures, where the same texture is used for repetitive parts of the building. We captured query images with cameras by walking around the cathedral, with the camera focused on the building. We constructed a MVS model from these images using the RealityCapture software and aligned the four models used for the experiments against this MVS model.

**Aachen.** We use a CAD model that contains buildings obtained by extruding building outlines based on human modeling from aerial images [33], as well as hand-modelled buildings. The model was kindly provided by the authors of [33] and was textured using aerial images. Viewing the model from the ground level thus leads to severe distortions due to the small viewing angle between the facades and the aerial images. We use the query images from the Aachen Day-Night v1.1 dataset [79, 80, 109]. The CAD model does not cover the full extent of the Aachen Day-Night v1.1 model. Thus, we only use the queries that depict (part of) the CAD model. The poses of the query images are not publicly available. Hence, we aligned the CAD model to a MVS model, reconstructed from the database images of the Aachen dataset using Colmap [82, 84], which are in the same coordinate system as the queries. This allows us to use the evaluation service at visuallocalization.net to measure pose accuracy for our experiments on Aachen.

### 4. Evaluation Protocol

This paper studies how the level of detail of 3D meshes downloaded from the Internet impacts visual localization performance. The goal is to answer the question to which degree such 3D models can be used to replace the classical scene representations, constructed from images, used in the literature. The following describes the localization system we use and how we measure localization performance.

**Visual localization.** We use the state-of-the-art MeshLoc approach [63] for our experiments. MeshLoc represents the scene via a 3D mesh and a set of database images with known poses. Image retrieval is used to find the top-k most similar database images for each query. 2D-2D matches between the query and the retrieved images are lifted to 2D-3D matches using the 3D mesh and the known poses of the database images. The camera pose of the query is then estimated from these 2D-3D matches using a P3P solver [65] inside RANSAC [25] with local optimization [51].

The 3D meshes used in [63] align well with the real scene geometry. [63] show that using renderings of the meshes instead of the original photos leads to comparable pose accuracy. They further show that even renderings of uncolored meshes can lead to accurate poses, as long as the meshes contain enough geometric detail. This observation motivated us to explore the more abstract (less geometrically detailed) CAD models available on the Internet.

In [63], the synthetic images are rendered from the poses of the original database images. However, we only have a 3D model of the scene and no database images. We thus use a simple approach to sampling camera poses around the model, from which we then render database images for each Internet model: We place cameras on the surfaces of multiple spheres with different radii, all centered around the center of gravity of the 3D model (cf. Fig. 5a). All cameras are looking at this center. The center and radii of the spheres, along with elevation limits and angular sampling periods, are manually adjusted to fit a particular model’s geometry.

**Performance measures.** A commonly used measure for localization performance computes the percentage of query images with camera pose estimates that are within given error thresholds of the ground truth [11, 43, 79, 87]. This requires that the scale of the scene is known, which is the case for the Aachen Day-Night v1.1 dataset [79, 80, 109] that is used to define the queries for the Aachen scene. By aligning the CAD model to the MVS mesh obtained from the database images of the Aachen dataset, we can use the ground truth poses from [79, 109] for evaluation.

However, this measure is not directly applicable to the other datasets, where the scale of the scene is not given. Rather than measuring errors in meters or degrees, we thus measure reprojection errors, which are independent of the scale of the scene. We use the Dense Correspondence Reprojection Error (DCRE) measure from [105]. Given a ground truth and an estimated pose, and a depth map of the 3D model from the view of the ground truth pose, the DCRE is measured as follows [105]: for each pixel in the depth map, we obtain a 3D point in the world coordinate system of the 3D model. Each 3D point is then projected into the image using the ground truth and the estimated pose, resulting in a set of 2D-2D correspondences. We measure the mean Euclidean distance between correspondences (mean DCRE) and the maximum Euclidean distance between correspondences (max. DCRE). Both DCRE variants measure how the change in pose between ground truth and estimate affects the quality of the alignment of the pose w.r.t. to the 3D model. Smaller DCREs correspond to better alignments and are a direct measure of pose quality for AR applications.

We consider two ways to define ground truth poses re-
required for measuring the DCREs: For each scene, we rigidly align the Internet models to a MVS model computed from the query images (the query model). For the alignment, we use ICP [72], starting from a manual initialization using 3D point correspondences. This alignment defines the poses of the query images w.r.t. the Internet models and we use these poses as one set of ground truth poses. We refer to these poses as global alignment (GA) poses.

DCREs computed w.r.t. to the GA poses measure how well the poses estimated via visual localization align with the poses from which the photos were taken in the scene. The underlying assumption is that the Internet model can be aligned well to the query model using a rigid transformation. However, the geometry of the query and the Internet models might differ, e.g., the ratios of the 3D model width and height might be different. Thus, a rigid alignment between the model types might be insufficient and the GA poses will not reflect the “optimal” pose of the query w.r.t. the Internet model. We thus consider a second set of ground truth poses, obtained by refining the GA poses w.r.t. each Internet model: given the GA pose and the depth map (generated by rendering the query model) for a query image, we align this depth map to the Internet model via ICP [72]. We will refer to these poses as locally refined (LR) poses.

5. Experimental Evaluation

We perform three sets of experiments: (1) we measure the level of geometric fidelity of the different models. This will later allow us to conclude how the level of appearance fidelity affects pose accuracy. (2) we measure localization accuracy via DCREs for the two definitions of ground truth poses, as well as pose accuracy in meters and degrees on the Aachen scene. (3) we isolate the impact of geometric fidelity on the localization process.

Measuring geometric fidelity. We measure how accurately the Internet models capture the real geometry of the scene. To this end, we compute distances in 3D between the query model and the Internet models, with the query model serving as an approximation to the true scene geometry. For

Figure 3. Cumulative histograms of distances from query model vertices to the nearest vertex from the Internet models. Smaller distances indicate a higher level of geometric fidelity for the Internet models.

Figure 4. Cumulative histograms of the mean DCRE over all query images in a scene for the ground truth poses obtained via global alignment (GA) and local refinement (LR). We show the DCRE as the percentage of the image diagonal.
each 3D point in the query model mesh, we find the nearest vertex in an Internet model mesh. We appropriately subdivide the Internet models to handle scenarios where an Internet model consists of only a few large polygons. The query model often shows only part of the Internet model, which is why we do not measure distances in the other direction.

Fig. 3 shows cumulative histograms over these distances. Here, smaller distances indicate a higher level of geometric fidelity. Note that for all scenes except Aachen, the 3D distances have no physical meaning. Within a scene, we can compare different models as they are all aligned to the query model and thus share the same scale. However, comparing models between scenes is not possible. For Aachen, due to the known scale of the Aachen Day-Night v1.1 [79,80,109] model, one model unit corresponds to one meter. As can be seen, the level of geometric fidelity of the Internet models used in our benchmark varies significantly.

Measuring localization performance via DCREs. We localize real images against the Internet models using the MeshLoc [65] pipeline. We evaluate different features and matchers inside MeshLoc: LoFTR [91], SuperGlue [75], and the combination of Patch2Pix [111] and SuperGlue [75]. The main paper only shows mean DCRE results. Max. DCRE results are shown in the supp. mat.

Fig. 4 shows cumulative distributions of the mean DCRE over all query images for a scene. Comparing the results with those shown in Fig. 3 and the visualization from Fig. 2, we can observe the following: (1) the best results are achieved by Internet models with both (relatively) high geometric and appearance level of detail: the Notre Dame A and B, and Pantheon A models are among the geometrically most accurate ones and provide high-quality textures. For all three, most images have a mean DCRE of 10% or less, which is comparable to what has been reported in the literature for scene representations created from RGB(D) images (cf. supp. mat. of [11]). This clearly shows that the approach of using models downloaded from the Internet as a scene representation for visual localization is feasible and can lead to high pose accuracy. (2) higher appearance fidelity can significantly compensate for a lower geometric fidelity. E.g., the Pantheon B and Reichstag B models are both significantly less accurate than other models from the same scene. Still, both models lead to very good localization performance. (3) if there are large discrepancies between the Internet model and the real world, then localization quickly fails. E.g., none of the models for the St. Vitus and St. Peter’s square scenes performs well. For the St. Vitus A and B models, this can be explained by the fact that the models were generated from drone imagery, leading to severe texture and model distortions. As a result, images taken close to the scene from the ground lead to localization failures (cf. Fig. 5b). Similarly, the St. Vitus C and D...
models lack the geometric and appearance detail necessary to localize such images. However, it is still possible to localize images taken relatively far from the scene, where less geometric and appearance detail is needed (cf. Fig. 5b). (4) we generally observe a higher performance with the global alignment (GA) ground truth poses compared to the locally refined (LR) poses. We attribute this to a lack of geometric detail in some of the models (in particular for the Reichstag scene), which allows the ICP algorithm [72] to significantly alter poses. (5) there can be significant differences in performance depending on the type of local features used, e.g., the Reichstag E and St. Peter’s square D models. Both these models do not have any texture, indicating that this is still a very challenging problem. Since such 3D models are often very compact in terms of memory footprint, being able to use them for localization is a problem of interest.

**Measuring via absolute errors.** Tab. 2 reports localization performance using the benchmark from [79, 96] for evaluation. The CAD model allows coarse localization despite its simplistic geometry and low-quality textures. Thus, this simple model is a viable alternative to a more detailed mesh if coarse pose accuracy is sufficient.

**Isolating the impact of geometric fidelity.** There are two main failure cases when localizing real images against Internet models: (1) Due to significant changes in appearance, insufficiently many matches are found. (2) Camera pose estimation computes a rigid alignment between the image and the Internet model. The pose estimation stage thus may fail if the geometry of the Internet model is too distorted. In this experiment, we isolate the impact of geometric fidelity on the localization. The Internet models are only used to lift the 2D-2D matches between the real images to 2D-3D matches that can be used for pose estimation. As can be seen from Fig. 6, the limiting factor often is the feature matching stage. Given enough feature matches, even geometrically less accurate models can produce relatively accurate poses (e.g., see results for St. Peter’s Square in Fig. 4 and Fig. 6). Thus, research on features and matchers that can handle this challenging task is a very promising direction for future work.

To better measure the influence of geometry inconsistency on the success of localization, we simulate different levels of distortion for one of the most precise models (Notre Dame A). We scale the model in a non-uniform way and use renderings of the resulting model for feature matching and pose estimation. Details on the scaling procedure are provided in the supp. mat. As can be seen in Fig. 7, increasing scale inconsistency significantly reduces localization success. The depth scaling does not influence the localization as much because all the query images were captured from the front facade side of the building, so only the depth changes of the more minor facade elements, such as doors, windows, or pillars, are visible. The results show that pose estimation strategies that actively account for non-uniform scaling are required for heavily distorted models.

We also investigate the influence of geometry and appearance simplification in Sec. 4 of the supp. mat.

**6. Conclusion**

In this paper, we have explored an alternative for scene representations in visual localization pipelines, i.e., 3D models that are readily available on the Internet. Such a representation is intriguing, as it, e.g., offers the promise of setting up AR experiences for users that are visiting places that the author creating the experience never visited. Yet, it also comes with many challenges due to imperfections of these 3D models. We studied how such imperfections affect localization accuracy. Detailed experiments show that 3D models from the Internet represent a promising new category of scene representations, which at the same time also open new directions for improvement in visual localization. In particular, research on matching real images against more abstract scene representation is an interesting direction for future work. To foster research on this interesting task, we publish the benchmark (v-pnk.github.io/cadloc).

**Acknowledgements.** This work was supported by the European Regional Development Fund under project IMPACT (No. CZ.02.1.01/0.0/0.0/15_003/0000468), a Meta Reality Labs research award under project call 'Benchmarking City-Scale 3D Map Making with Mapillary Metropolis', the Grant Agency of the Czech Technical University in Prague (No. SGS21/119/OHK3/2T/13), the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 “Research Center for Informatics”, and the Czech Science Foundation (GAČR) EXPRO Grant No. 23-07973X.

<table>
<thead>
<tr>
<th>method</th>
<th>day</th>
<th>night</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAD - LoFTR [91]</td>
<td>0.4 / 6.8 / 71.3</td>
<td>0.6 / 7.0 / 72.5</td>
</tr>
<tr>
<td>CAD - Patch2Pix [111] + SuperGlue [75]</td>
<td>0.7 / 5.6 / 73.0</td>
<td>1.1 / 7.6 / 74.3</td>
</tr>
<tr>
<td>CAD - SuperGlue [73]</td>
<td>0.6 / 6.1 / 72.2</td>
<td>1.8 / 8.8 / 73.3</td>
</tr>
</tbody>
</table>

Table 2. Localization results for the Aachen scene for different types of local features. We report the % of query images localized within (0.25m, 2°), (0.5m, 5°), (5m, 10°) of the ground truth pose. We report results only for queries that observe the CAD model.
References

[17] Tommaso Cavallari, Luca Bertinetto, Jishnu Mukhoti, Philip Torr, and Stuart Golodetz. Let’s take this online: Adapting scene coordinate regression network predictions for online RGB-D camera relocalisation. In 3DV, 2019. 1, 2
D. Sibbing, T. Sattler, B. Leibe, and L. Kobbelt. SIFT-Realistic Rendering. In JDV, 2013. 1, 2


Hajime Taira, Ignacio Rocco, Jiri Sedlar, Masatoshi Okutomi, Josef Sivic, Tomas Pajdla, Torsten Sattler, and Akihiko Torii. Is This the Right Place? Geometric-Semantic Pose Verification for Indoor Visual Localization. In The IEEE International Conference on Computer Vision (ICCV), 2019. 2

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl S. Ni, Douglas N. Poland, Damian Borth, and Li-Jia Li. YFCC100M: the new data in multimedia research. Commun. ACM, 59:64–73, 2016. 4


Akihiko Torii, Relja Arandjelović, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In CVPR, 2015. 2


Florian Walch, Caner Hazirbas, Laura Leal-Taixé, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-Based Localization Using LSTMs for Structured Feature Correlation. In ICCV, 2017. 1, 2


Bernhard Zeisl, Torsten Sattler, and Marc Pollefeys. Camera pose voting for large-scale image-based localization. In ICCV, 2015. 1, 2


Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Reference Pose Generation for Long-term Visual Localization via Learned Features and View Synthesis. IJCV, 2020. 1, 2, 3, 4, 5, 7
