Semantic Match Consistency for Long-Term Visual Localization

Carl Toft[1], Erik Stenborg[1], Lars Hammarstrand[1], Lucas Brynte[1], Marc Pollefeys[2,3], Torsten Sattler[2], Fredrik Kahl[1]

[1]Department of Electrical Engineering, Chalmers University of Technology, Sweden
[2]Department of Computer Science, ETH Zürich, Switzerland
[3]Microsoft, Switzerland

Abstract. Robust and accurate visual localization across large appearance variations due to changes in time of day, seasons, or changes of the environment is a challenging problem which is of importance to application areas such as navigation of autonomous robots. Traditional feature-based methods often struggle in these conditions due to the significant number of erroneous matches between the image and the 3D model. In this paper, we present a method for scoring the individual correspondences by exploiting semantic information about the query image and the scene. In this way, erroneous correspondences tend to get a low semantic consistency score, whereas correct correspondences tend to get a high score. By incorporating this information in a standard localization pipeline, we show that the localization performance can be significantly improved compared to the state-of-the-art, as evaluated on two challenging long-term localization benchmarks.

Keywords: Visual localization, semantic segmentation, camera pose estimation, outlier rejection, self-driving cars

1 Introduction

Visual localization, i.e., estimating the camera pose of a query image with respect to a scene model, is one of the core problems in computer vision. It plays a central role in a wide range of practical applications, such as Structure-from-Motion (SfM) [43], augmented reality [9], and robotics [31], where visual navigation for autonomous vehicles has recently been receiving considerable attention.

Traditional approaches to the visual localization problem [27–29,38,40,47,58] rely on local feature descriptors to establish correspondences between 2D features found in a query image and 3D points in an SfM model of the scene. These 2D-3D matches are then used to estimate the camera pose of the query image by applying an n-point-pose solver, e.g., [23], inside a RANSAC loop [18]. Learning-based alternatives exist [6,7,21,22,51], but are either less accurate than feature-based approaches [40,51] or struggle to handle larger scenes [7,21,37]. Feature-based approaches thus still represent the current state-of-the-art [37,40,51].

Existing feature-based methods for visual localization tend to work very well when the query image is taken under similar conditions as the database images

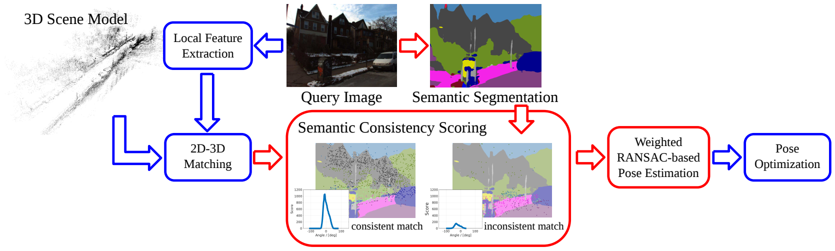

Fig. 1. Illustration of the visual localization approach proposed in this paper. We extend the standard localization pipeline (blue boxes) to include a semantic consistency score (red boxes). Our approach rates the consistency of each 2D-3D match and uses the score to prioritize more consistent matches during RANSAC-based pose estimation.

used for creating the 3D model. However, feature matching performance suffers if the localization and mapping stages occur far apart in time [37], e.g., in different weather conditions, between day and night, or across different seasons. As feature detectors become less repeatable and feature descriptors less similar, localization pipelines struggle to find enough correct 2D-3D matches to facilitate successful pose estimation. One possible solution is to map the scene in as wide a range of different conditions as possible. Yet, 3D model construction and extensive data collection are costly, time-consuming, and tedious processes. At the same time, the resulting models consume a significant amount of memory. Developing localization algorithms that work well across a wide range of conditions, even if the 3D model is constructed using only a single condition, is thus desirable.

This paper presents a step towards robust algorithms for long-term visual localization through a novel strategy for robust inlier / outlier detection. The main insight is that semantic information can be used as a weak supervisory signal to distinguish between correct and incorrect correspondences: Given a semantic segmentation for each database image, we can assign a semantic label to each 3D point in the SfM model. Given a pose estimate for a query image, we can project the 3D points into a semantic segmentation of the query image. An estimate close to the correct pose should lead to a *semantically consistent* projection, where each point is projected to an image region with the same semantic label. Based on this idea, we assign each 2D-3D match a semantic consistency score, where high scores are assigned to matches more likely to be correct. We later use these scores to bias sampling during RANSAC-based pose estimation. See Fig. 1 for an overview. While conceptually simple, this strategy leads to dramatic improvements in terms of localization rate and pose accuracy in the long-term localization scenario. The reason is that, unlike existing methods, our approach takes advantage of unmatched 3D points, and consequently copes much better with situations in which only few correct matches can be found.

While the idea of using semantics for localization is not new, cf. [42, 48], the challenge is to develop a computationally tractable framework that takes advantage of the available information. In detail, this paper makes the following

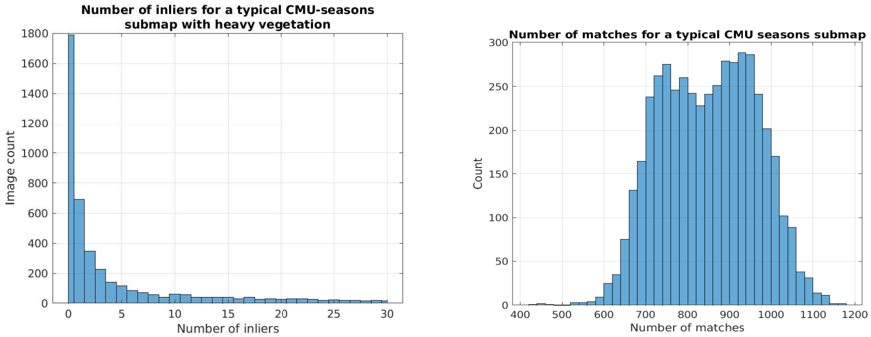

Fig. 2. For each ground truth camera pose, we have counted how many true inlier 2D-3D matches have been obtained at the matching stage (shown in the left figure). Note that in around 70% of the cases, there are less than 7 true inliers. For structure-based methods, typically 12 or more consistent 2D-3D matches are required to compute and verify a camera pose, cf. [27,28,36]. The right figure shows a histogram over the number of matches for the same submap from the CMU Seasons dataset [5,37].

contributions: 1) We present a new localization method that incorporates both standard feature matching and semantic information in a robust and efficient manner. At the center of our method is a novel semantic consistency check that allows us to rate the quality of individual 2D-3D matches. 2) We extensively evaluate and compare our method to the current state-of-the-art on two benchmarks for long-term visual localization. Our experimental results show significant improvements by incorporating semantics, particularly for challenging scenarios due to change of weather, seasonal, and lighting conditions.

The remainder of this paper is structured as follows: Sec. 2 reviews related work. Sec. 3 derives our semantic consistency score and shows how it can be incorporated into a state-of-the-art localization pipeline. Sec. 4 extensively evaluates our approach in the context of long-term localization.

2 Related Work

Traditionally, there are two approaches to visual localization: The first one uses image retrieval techniques to find the most relevant database images for each query image [1,11,24,39,41,49,56,57]. The pose of the query image is then approximated by the pose of the top-ranked retrieved image [11] or computed from the top-k ranking database images [40,56,59]. Instead of explicitly representing a scene by a database of images, another approach is to implicitly model a scene by a CNN trained for pose regression [21,22,51] or place classification [52].

The second approach is based on 3D scene models, typically reconstructed using SfM. Such *structure-based* methods assign one or more feature descriptors, e.g., SIFT [30] or LIFT [53], to each 3D point. For a given query image, 2D-3D correspondences are established using descriptor matching. These matches are then used to estimate the camera pose. Compared to image-retrieval approaches, structure-based methods tend to provide more accurate camera poses [40]. Yet, it is necessary that enough correct matches are found to not only estimate a pose, but also to verify that the pose is indeed correct, e.g., through inlier counting. As shown in Fig. 2 and [37], these conditions are often not satisfied when the query images are taken under significantly different conditions compared to the database images. Our approach extends structure-based methods by incorporating semantic scene understanding into the pose estimation stage.

Structure-based approaches for visual localization can be classified based on their efficiency and ability to handle more complex scenes. Approaches based on prioritized matching [12, 28, 36] focus on efficiency by terminating correspondence search once a fixed number of matches has been found. In order to handle more complex environments, robust structure-based approaches either relax the matching criteria [8, 27, 38, 47, 58] or restrict the search space [20, 27, 29, 38]. The latter type of methods use image retrieval [20, 38] or co-visibility information [27, 29] to determine which parts of the scene are visible in a query image, potentially allowing them to disambiguate matches. The former type handles the larger amount of outliers resulting from a more relaxed matching stage through deterministic outlier filtering. To this end, they use geometric reasoning to determine how consistent each match is with all other matches [8, 47, 58]. Especially when the gravity direction is known, which is typically the case in practice (e.g., via sensors or vanishing points), such approaches can handle outlier ratios of 99% or more [47, 58]. Our approach combines techniques from geometric outlier filtering [47, 58] with reasoning based on scene semantics. This enables our method to better handle scenarios where it is hard to find correct 2D-3D matches.

An alternative to obtaining 2D-3D correspondences via explicit feature matching is to directly learn the matching function [6, 7, 10, 33, 45, 50]. Such methods implicitly represent the 3D scene structure via a random forest or CNN that predicts a 3D scene coordinate for a given image patch [33]. While these methods can achieve a higher pose accuracy than feature-based approaches [7], they also have problems handling larger outdoor scenes to the extent that training might fail completely [7, 37, 42].

The idea of using semantic scene understanding as part of the visual localization process has gained popularity over the last few years. A common strategy is to include semantics in the matching stage of visual localization pipelines, either by detecting and matching objects [3, 4, 15, 35, 44, 55] or by enhancing classical feature descriptors [2, 25, 46]. The latter type of approaches still mainly relies on the strength of the original descriptor as semantics only provide a weak additional signal. Thus, these approaches do not solve the problem of finding enough correct correspondences, which motivates our work. Recent work shows that directly learning a descriptor that encodes both 3D scene geometry and semantic information significantly improves matching performance [42]. Yet, this approach requires depth maps for each query image, e.g., from stereo, which are not necessarily available in the scenario we are considering.

In contrast to the previously discussed approaches, which aim at improving the matching stage in visual localization, our method focuses on the subsequent pose estimation stage. As such, most similar to ours is existing work on semantic hypothesis verification [14] and semantic pose refinement [48]. Given a hypothesis for the alignment of two SfM models, Cohen et al. [14] project the 3D points of one model into semantic segmentations of the images used to reconstruct the other model. They count the number of 3D points projecting into regions labelled as "sky" and select the alignment hypotheses with lowest number of such free-space violations. While Cohen et al. make hard decisions, our approach avoids them by converting our semantic consistency score into sampling probabilities for RANSAC. Our approach aims at improving pose hypothesis generation while Cohen et al. only rate given hypotheses. Given an initial camera pose hypothesis, Toft et al. [48] use semantics to obtain a refined pose estimate by improving the semantic consistency of projected 3D points and curve segments. Their approach could be used as a post-processing step for the poses estimated by our method.

3 Semantic Match Consistency for Visual Localization

As outlined above, long-term localization is a hard problem due to the difficulty of establishing reliable correspondences. Our approach follows a standard feature-based pipeline and is illustrated in Fig. 1. Our central contribution is a novel semantic consistency score that is used to determine which matches are likely to be correct. Building on top of existing work on geometric outlier filtering [47,58], we generate a set of pose hypotheses for each 2D-3D correspondence established during the descriptor matching stage. These poses are then used to measure the semantic consistency of each match. We then use the consistency scores to bias sampling inside RANSAC towards semantically consistent matches, allowing RANSAC to focus more on matches more likely to be correct.

Specifically, for each pose hypothesis generated by a given 2D-3D match, we project the visible 3D structure into the corresponding camera. Since each 3D point is endowed with a semantic label, it is possible to compare the observed semantic label in the query image with the label assigned to the 3D point. The semantic inlier count for that pose is given by the number of 3D points that project into pixels whose semantic class agrees with that of the point. The semantic consistency for the 2D-3D correspondence is then defined as the maximum semantic inlier count over all hypotheses generated by that correspondence.

Our approach offers a clear advantage over existing outlier filtering strategies [8,47,58]: Rather than being restricted to the 2D-3D correspondences found during the matching stage, the semantic consistency score allows us to also use unmatched 3D points when rating the 2D-3D matches. As a result, our approach is better able to handle scenarios in which it is hard to find many correct matches.

In this section, we present our proposed localization method based on semantic consistency in detail. We first introduce necessary notation. Sec. 3.1 explains the pose hypothesis generation stage. Our semantic consistency score is then described in Sec. 3.2. Finally, Sec. 3.3 summarizes the complete pipeline.

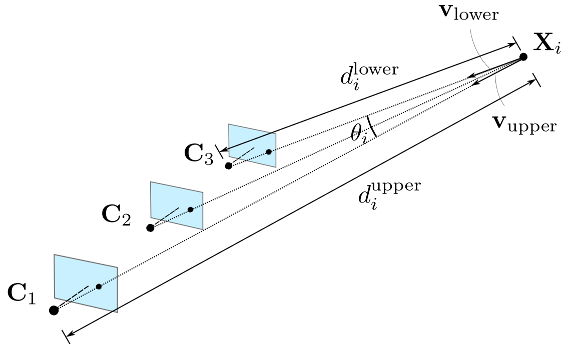

Fig. 3. Example triangulation of a point during the 3D reconstruction of the point cloud. The example shows a 3D point triangulated from three observations. The quantities θ_i, d_i^{lower} and d_i^{upper} as defined in the text are shown. The vector $\boldsymbol{v}_i$ for this point is the unit vector in the middle between $\boldsymbol{v}^{\text{lower}}$ and $\boldsymbol{v}^{\text{upper}}$.

Notation. We compute the camera pose of a query image relative to a 3D point cloud that has been pre-computed using a regular Structure-from-Motion pipeline. The 3D map is defined as a set of 3D points

$$\mathcal{M} = \{(\boldsymbol{X}_i, c_i, \boldsymbol{f}_i, \boldsymbol{v}_i, \theta_i, d_i^{\text{lower}}, d_i^{\text{upper}})\}_{i=1}^{N} \ , \tag{1}$$

where N is the number of 3D points in the model. Each 3D point is defined by its 3D coordinates $\boldsymbol{X}_i$, its class label c_i (e.g., vegetation, road, *etc*), visibility information, and its corresponding (mean) feature descriptor $\boldsymbol{f}_i$. We encode the visibility information of a point as follows (cf. Fig. 3): $\boldsymbol{v}_i$ is a unit vector pointing from the 3D point towards the mean direction from which the 3D point was seen during reconstruction. It is computed by determining the two most extreme viewpoints from which the point was triangulated ($\boldsymbol{v}_{\text{lower}}$ and $\boldsymbol{v}_{\text{upper}}$ in the figure) and choosing the direction half-way between them. The angle θ_i is the angle between the two vectors. The quantities d_i^{lower} and d_i^{upper} denote the minimum and maximum distances, respectively, from which the 3D point was observed during SfM. Note that all this information is readily provided by SfM.

The semantic class labels are found by performing a pixelwise semantic labelling of all database images. For a 3D point in the SfM model, we assign its label c_i to the semantic class it was most frequently observed in.

3.1 Generating Camera Pose Hypotheses

In order to determine the semantic consistency of a single 2D-3D match $\boldsymbol{x}_j \leftrightarrow \boldsymbol{X}_j$, we compute a set of plausible camera poses for this match. We thereby follow the setup used by geometric outlier filters [47, 58] and assume that the gravity direction $\boldsymbol{g}$ in the local camera coordinates of the query image is known. This assumption is not restrictive as the gravity direction can typically be estimated very reliably from sensors or from vanishing points. In the experiments, the

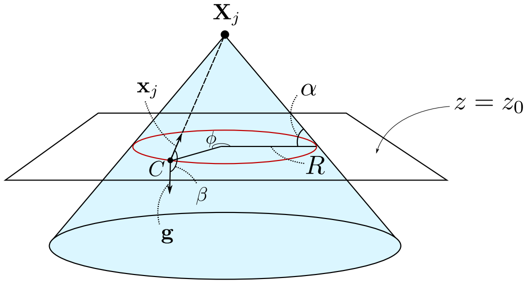

Fig. 4. If the gravity direction is known in the camera's coordinate system, a single 2D-3D match $\boldsymbol{x}_j \leftrightarrow \boldsymbol{X}_j$ constrains the camera center to lie on a cone with $\boldsymbol{X}_j$ at the vertex, and whose axis is parallel to the gravity direction. If the camera height is also known, the camera center must lie on a circle (shown in red).

gravity direction in local camera coordinates was extracted from the ground truth camera pose.

Knowing the intrinsic camera calibration and the point position $\boldsymbol{X}_j$, the correspondence can be used to restrict the set of plausible camera poses under which $\boldsymbol{X}_j$ exactly projects to $\boldsymbol{x}_j$ [47,58]: The camera center must lie on a circular cone with $\boldsymbol{X}_j$ at its vertex (cf. Fig. 4). To see this, let the position of the camera center be $\boldsymbol{C}$ and the coordinates of $\boldsymbol{X}_j$ be $(x_j, y_j, z_j)^T$. In a slight abuse of notation, let $\boldsymbol{x}_j$ be the normalized viewing direction corresponding to the matching 2D feature. Since the gravity vector $\boldsymbol{g}$ in local camera coordinates is known, we can measure the angle β between the gravity direction and the line that joins $\boldsymbol{C}$ and $\boldsymbol{X}_j$ as $\beta = \arccos(\boldsymbol{g}^T \boldsymbol{x}_j)$. Assuming that the gravity direction in the 3D model coincides with the z-axis, the angle between the line joining C and $\boldsymbol{X}_j$ and the $xy-$plane then is

$$\alpha = \arccos(\boldsymbol{g}^T \boldsymbol{x}_j) - \pi/2 \ . \tag{2}$$

The set of points $\boldsymbol{C}$ such that the angle between the $xy-$plane and the line joining $\boldsymbol{C}$ and $\boldsymbol{X}_j$ equals α is a cone with $\boldsymbol{X}_j$ as the vertex. Note that the cone's position and opening angle are fully determined by $\boldsymbol{g}$ and the correspondence $\boldsymbol{x}_j \leftrightarrow \boldsymbol{X}_j$. Also note that the camera rotation is fully determined at each point of the cone [58]: two of the rotational degrees of freedom are fixed by the known gravity direction and the last degree is fixed by requiring that the viewing direction of $\boldsymbol{x}_j$ points to $\boldsymbol{X}_j$. As a result, two degrees-of-freedom remain for the camera pose, corresponding to a position on the cone's surface.

Often, the camera height z_0 can be roughly estimated from the typical depth of the 3D point in the SfM model[1]. Knowing the camera height removes one degree of freedom. As a result, the camera must lie on the circle with radius R given by $R = |z_j - z_0|/|\tan\alpha|$, which lies in the plane $z = z_0$, and whose center

[1] This strategy is used in the experiments to estimate the camera heights.

Algorithm 1 Semantic consistency score calculation for single correspondence

```
procedure CalculateScore($\boldsymbol{x}_j, \boldsymbol{X}_j, \boldsymbol{g}, z_0, \mathcal{M}$)
    maxScore ← 0
    $\alpha \leftarrow \arccos(\boldsymbol{g}^T \boldsymbol{x}_j) - \pi/2$    ▷ Angle of cone
    $R \leftarrow |z_j - z_0|/|\tan\alpha|$    ▷ Radius of circle of possible camera poses
    for $\phi \in \{0°, 1°, \ldots, 359°\}$ do
        score ← 0
        Calculate camera center $\boldsymbol{C}(\phi)$ using $\phi$
        Calculate projection matrix $P(\phi)$ using $R$, $\boldsymbol{C}(\phi)$
        for $\boldsymbol{X}_k \in \mathcal{M}$ do
            $\boldsymbol{u} \leftarrow P(\phi)\boldsymbol{X}_k$    ▷ Project 3D point into query image
            if $\boldsymbol{C}(\phi) \in \mathcal{V}_k$ and $I_{\text{semantic}}(\boldsymbol{u}) = c_k$ then
                score ← score + 1    ▷ Point is visible and semantically consistent
        if score > maxScore then
            maxScore ← score
            $\boldsymbol{C}_{\text{best}} \leftarrow \boldsymbol{C}(\phi)$
    return (maxScore, $\boldsymbol{C}_{\text{best}}$)
```

point is the point (x_j, y_j, z_0) [58]. For a single correspondence $\boldsymbol{x}_j \leftrightarrow \boldsymbol{X}_j$, we thus generate a set of plausible camera poses by varying an angle ϕ that defines positions on this circle (cf. Fig. 4).

3.2 Measuring Semantic Match Consistency

Given a 2D-3D match $\boldsymbol{x}_j \leftrightarrow \boldsymbol{X}_j$ and its corresponding set of camera pose hypotheses (obtained by discretizing the circle into evenly spaced points), we next compute a *semantic consistency score* as described in Alg. 1.

For a camera pose hypothesis corresponding to an angle ϕ, we project the semantically labelled 3D points from the SfM model into a semantic segmentation of the query image. We then count the number of 3D points that project to a pixel whose semantic class matches that of the 3D point. For each pose on the circle, we thus find the number of 3D points that agree with the semantic labelling of the query image. The semantic consistency score for a match $\boldsymbol{x}_j \leftrightarrow \boldsymbol{X}_j$ is then defined as the maximum number of semantic inliers while sweeping the angle ϕ. Note that we project all 3D points in the model, not only the correspondences found via descriptor matching. This means that the calculation of the consistency score is not dependent of the quality of the correspondences.

Since we are using all 3D points in a model, we need to explicitly handle occlusions: a 3D point is not necessarily visible in the image even though it projects inside the image area for a given camera pose. We do so by defining a visibility volume for each 3D point from the corresponding visibility information $\boldsymbol{v}_i$, θ_i, d_i^{lower} and d_i^{upper}. The volume for the i^{th} point is defined as

$$\mathcal{V}_i = \left\{\boldsymbol{X} \in \mathbf{R}^3 : d_i^{\text{lower}} < ||\boldsymbol{X} - \boldsymbol{X}_i|| < d_i^{\text{upper}}||, \angle(\boldsymbol{X} - \boldsymbol{X}_i, \boldsymbol{v}_i) < \theta_i\right\} \quad . \tag{3}$$

A 3D point is only considered visible from a camera pose with with its center at $\boldsymbol{C}$ if $\boldsymbol{C} \in \mathcal{V}_i$. The intuition is that a 3D point only contributes to the semantic score if it is viewed from approximately the same distance and direction as the 3D point was seen from when it was triangulated during SfM. This is not too much of a restriction since local features are not completely invariant to changes

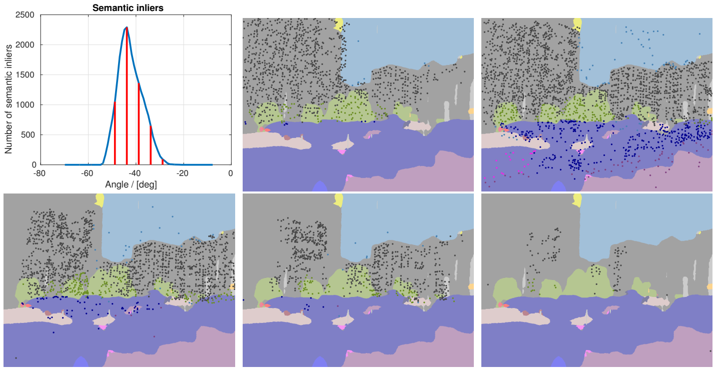

Fig. 5. An example for sweeping over the angle ϕ for a single correct 2D-3D match. The upper left figure shows the number of semantic inliers as a function of ϕ. The five images with projected points correspond to the red lines in the top left and are shown in order from left to right. The upper right image corresponds to the angle that yielded the largest number of semantic inliers.

in viewpoint, i.e., features naturally do not match anymore if a query image is taken too far away from the closest database image.

To further speed up the semantic scoring, we limit the set of labelled points that are projected into the image. For a 2D-3D match, only those 3D points inside a cylinder with radius R whose axis aligns with the gravity direction and goes through the 3D point $\boldsymbol{X}_j$ are considered.

Discussion. Intuitively, if a match $\boldsymbol{x}_j \leftrightarrow \boldsymbol{X}_j$ is correct, we expect the number of semantic inliers to be large for values of ϕ that correspond to camera poses close to the ground truth pose, and small for values of ϕ that yield poses distant to the ground truth pose. An example of this behavior is shown in Fig. 5. On the other hand, if a match is an outlier, we would expect only a small number of semantic inliers for all values of ϕ (cf. Fig. 1).

Naturally, the distribution of the number of semantic inliers over the angle ϕ and the absolute value of the semantic consistency score depend on how "semantically interesting" a scene is. As shown in Fig. 2, the case where many different classes are observed leads to a clear and high peak in the distribution. If only a single class is visible, e.g., "building", we can expect a more uniform distribution, both for correct and incorrect matches. As shown later, our approach degenerates to the standard localization pipeline in this scenario.

3.3 Full Localization Pipeline

Fig. 1 shows our full localization pipeline: Given a query image, we extract local (SIFT [30]) features and compute its semantic segmentation. Using approximate nearest neighbor search, we compute 2D-3D matches between the query features and the 3D model points. We follow common practice and use Lowe's ratio test to filter out ambiguous matches [30]. Similar to work on geometric outlier filtering [8, 47, 58], we use a rather relaxed threshold of 0.9 for the ratio test to avoid rejecting correct matches. Next, we apply our proposed approach to compute a semantic consistency score per 2D-3D match (cf. Alg. 1). For each correspondence, an estimate of the camera height z_0 is obtained by checking where the database trajectory (whose poses and camera heights are available) intersects the cone of possible poses. Lastly, we apply an n-point-pose solver inside a RANSAC loop for pose estimation, using 10'000 iterations.

We use the consistency scores to adapt RANSAC's sampling scheme. More precisely, we normalize each score by the sum of the scores of all matches. We interpret this normalized score as a probability p_j and use it to bias RANSAC's sampling, i.e., RANSAC selects a match $\boldsymbol{x}_j \leftrightarrow \boldsymbol{X}_j$ with probability p_j. This can thus be seen as a "soft" version of outlier rejection: instead of explicitly removing correspondences that seem to be outliers, it just becomes unlikely that they are sampled inside RANSAC. This strategy guarantees that our approach gracefully degenerates to a standard pipeline in semantically ambiguous scenes.

4 Experimental Evaluation

In this section we present experimental evaluations of the proposed algorithm on two challenging benchmark datasets for long-term visual localization. The datasets used are the *CMU Seasons* and *RobotCar Seasons* datasets from [37].

CMU Seasons. The dataset is based on the CMU Visual Localization dataset [5]. It consists of 7,159 database images that can be used for map building, and 75,335 query images for evaluation. The images are collected from two sideways facing cameras mounted on a car while traversing the same route in Pittsburgh on 12 different occasions over the course of a year. It captures many different environmental conditions, including overcast weather, direct sunlight, snow, and trees with and without foliage. The route contains urban and suburban areas, as well as parks mostly dominated by vegetation. All images are accompanied by accurate six degrees-of-freedom ground truth poses [37].

CMU Seasons is a very challenging dataset due to the large variations in appearance of the environment over time. Especially challenging are the areas dominated by vegetation, since these regions change drastically in appearance under different lighting conditions and in different seasons.

We used the Dilation10 network [54] trained on the Cityscapes dataset [16] to obtain the semantic segmentations. The classes used to label the 3D points were: sky, building, vegetation, road, sidewalk, pole and terrain/grass.

RobotCar Seasons. The dataset is based on a subset of the Oxford RobotCar dataset [32]. It was collected using a car-mounted camera rig consisting of 3 cameras facing to the left, right and rear of the car. The dataset consists of 26,121 database images taken at 8,707 positions, and 11,934 query images captured at 3,978 positions. All images are from a mostly urban setting in Oxford, UK, but they cover a wide variety of environmental conditions, including varying light

Table 1. Ablation study of our approach on the CMU Seasons dataset

Method / Setting m deg	Urban 0.25 / 0.5 / 5 2 / 5 / 10	Suburban 0.25 / 0.5 / 5 2 / 5 / 10	Park 0.25 / 0.5 / 5 2 / 5 / 10
Weighted, P3P	75.2 / 82.1 / 87.7	44.6 / 53.9 / 63.5	30.4 / 37.8 / 48.0
Unweighted, P3P	42.5 / 50.0 / 64.5	11.5 / 16.8 / 30.1	9.3 / 13.1 / 24.2
Weighted, P2P	**81.7** / **88.0** / **92.3**	**55.4** / **65.5** / **73.1**	**39.5** / **47.5** / **58.2**
Unweighted, P2P	72.9 / 80.0 / 87.0	41.3 / 50.8 / 61.8	29.7 / 37.0 / 49.1

Table 2. Ablation study of our approach on the RobotCar Seasons dataset

Method / Setting m deg	all day 0.25 / 0.5 / 5 2 / 5 / 10	all night 0.25 / 0.5 / 5 2 / 5 / 10
Weighted, P3P	**50.6** / **79.8** / **95.1**	7.6 / 21.5 / 45.4
Unweighted, P3P	47.0 / 74.8 / 91.9	0.5 / 4.2 / 16.0
Weighted, P2P	35.4 / 73.2 / 93.4	**13.0** / **34.1** / **63.1**
Unweighted, P2P	34.1 / 71.3 / 93.3	5.1 / 20.8 / 46.8

conditions at day and night, seasonal changes from summer to winter, and various weather conditions such as sun, rain, and snow. All images have a reference pose associated with them. The average reference pose error is estimated to be below 0.10m in position and 0.5° in orientation [37].

The most challenging images of this dataset are the night images. They both exhibit a big change in lighting, but also, due to longer exposure times, contain significant motion blur.

For the RobotCar dataset, semantic segmentations were obtained using the PSPNet network [60], trained jointly on the Cityscapes [16] and Mapillary Vistas [34] datasets[2]. Additionally, 69 daytime and 13 nighttime images from the RobotCar dataset [32] were manually annotated by us, and incorporated into the training, in order to alleviate generalization issues. The classes used to label the 3D points were: sky, building, vegetation, road, sidewalk, pole and terrain/grass.

Evaluation protocol. We follow the evaluation protocol from [37], i.e., we report the percentage of query images for which the estimated pose differs by at most Xm and Y° from their ground truth pose. As in [37], we use three different threshold combinations, namely (0.25m, 2°), (0.5m, 5°), and (5m, 10°).

4.1 Ablation Study

In this section we present an ablation study of our approach on both datasets.

The baseline is a standard, unweighted RANSAC procedure that samples each 2D-3D match with the same probability. We combine this RANSAC variant, denoted as *unweighted*, with two pose solvers: the first is a standard 3-point solver [19] (P3P) since the intrinsic calibration is known for all query images. The second solver is a 2-point solver [26] (P2P) that uses the known gravity direction.

[2] Starting from a network pretrained on Cityscapes, joint training was carried out by regarding 4 Cityscapes samples, 4 Mapillary Vistas samples and 1 RobotCar sample in each iteration. Mapillary Vistas labels were mapped to Cityscapes labels by us.

Table 3. Comparison of our approach, using semantic consistency scoring and the P3P pose solver, with state-of-the-art approaches on the CMU Seasons dataset

Method / Setting	Urban	Suburban	Park
m	0.25 / 0.5 / 5	0.25 / 0.5 / 5	0.25 / 0.5 / 5
deg	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10
ActiveSearch [36]	55.2 / 60.3 / 65.1	20.7 / 25.9 / 29.9	12.7 / 16.3 / 20.8
CSL [47]	36.7 / 42.0 / 53.1	8.6 / 11.7 / 21.1	7.0 / 9.6 / 17.0
DenseVLAD [49]	22.2 / 48.7 / 92.8	9.6 / 26.6 / **85.2**	10.3 / 27.0 / **77.0**
NetVLAD [1]	17.4 / 40.3 / **93.2**	7.7 / 20.1 / 80.5	5.6 / 15.7 / 65.8
PROSAC P3P [13]	56.7 / 64.0 / 74.2	30.6 / 38.3 / 49.1	20.0 / 25.4 / 35.1
Single-match P3P	59.3 / 66.8 / 76.2	24.6 / 32.4 / 44.6	16.8 / 22.2 / 32.6
Sem. rank. (**ours**)	**75.2** / **82.1** / 87.7	**44.6** / **53.9** / 63.5	**30.4** / **37.8** / 48.0

Table 4. Comparison of our approach, using semantic consistency scoring and the P3P pose solver, with state-of-the-art approaches on the RobotCar Seasons dataset

Method / Setting	all day	all night
m	0.25 / 0.5 / 5	0.25 / 0.5 / 5
deg	2 / 5 / 10	2 / 5 / 10
ActiveSearch [36]	35.6 / 67.9 / 90.4	0.9 / 2.1 / 4.3
CSL [47]	45.3 / 73.5 / 90.1	0.6 / 2.6 / 7.2
PROSAC P3P [13]	50.4 / 79.1 / 96.4	3.9 / 14.1 / 34.4
Single-match P3P	**50.7** / 79.3 / **97.2**	2.5 / 6.4 / 16.7
Sem. rank. (**ours**)	50.6 / **79.8** / 95.1	**7.6** / **21.5** / **45.4**

We compare both baselines against our proposed *weighted* RANSAC variant that uses our semantic consistency scores to estimate a sampling probability for each 2D-3D match. Again, we combine our approach with both pose solvers.

Tables 1 and 2 show the results of our ablation study. As can be seen, using our proposed semantic consistency scores (*weighted*) leads to clear and significant improvements in localization performance for all scenes and solvers on the CMU dataset. On the RobotCar dataset, we similar observe a significant improvement when measuring semantic consistency, with the exception of using the P2P solver during daytime. Interestingly, the P2P solver outperforms the P3P solver on both datasets using both RANSAC variants, with the exception of the daytime query images of the Oxford RobotCar dataset. The reason for this is likely due to sensitivity of the P2P solver to small noise in the ground truth vertical direction.

4.2 Comparison with State-of-the-Art

After demonstrating the benefit of our proposed semantic consistency scoring, we compare our localization pipeline against state-of-the-art approaches on both datasets, using the results reported in [37]. More concretely, we compare against ActiveSearch (AS) [36] and the City-Scale Localization (CSL) [47] methods, which represent the state-of-the-art in efficient and scalable localization, respectively. In addition, we compare against two image retrieval-based baselines, namely DenseVLAD [49] and NetVLAD [1], when their results are available in [37]. We omitted results for the methods LocalSfM, DenseSfM, ActiveSearch+Generalized Camera, and FABMAP [17] present in [37], since these use either a sequence of images (the latter two), costly SfM approaches coupled with a strong location prior (the former two), or use ground truth information (the former three), and are thus not directly comparable. For a fair comparison with AS and CSL, we use the variant of our localization pipeline that uses semantic consistency scoring and the P3P solver.

Tables 3 and 4 show the results of our comparison. As can be seen, our approach significantly outperforms both AS and CSL, especially in the high-precision regime. Especially the comparison with CSL is interesting as our pose generation stage is based on its geometric outlier filtering strategy. The clear improvements over CSL validate our idea of incorporating scene semantics into the pose estimation stage in general and the idea of using non-matching 3D points to score matches in particular.

On the CMU dataset, both DenseVLAD and NetVLAD can localize more query images in the coarse-precision regime (5 m, 10°). Both approaches represent images using a compact image-level descriptor and approximate the pose of the query image using the pose of the top-retrieved database image. Both methods do not use any feature matching between images. As shown in Fig. 6, this allows DenseVLAD and NetVLAD to handle scenarios with very strong appearance changes in which feature matching completely fails. Note that both DenseVLAD or NetVLAD could be used as a fallback option for our approach.

Interestingly, the P3P RANSAC baseline outperforms AS and CSL in several instances. This is likely due to differing feature matching strategies and different numbers of RANSAC iterations. Active Search uses a very strict ratio test, which causes problems in challenging scenes. CSL was evaluated on CMU Seasons by keeping all detected features (no ratio test), resulting in several thousand matches per image. CSL may have yielded better results with a ratio test.

In addition, we also compare our approach to two methods based on P3P RANSAC. The first is PROSAC [13], a RANSAC variant that uses a deterministic sampling strategy, where correspondences deemed more likely to be correct are given higher priority during sampling. In our experiments, the quality measure used was the Euclidean distance between the descriptors of the observed 2D point and the corresponding matched 3D point.

The second RANSAC variant employs a very simple single-match semantic outlier rejection strategy: all 2D-3D matches for which the semantic labels of the 2D feature and 3D point do not match are discarded before pose estimation.

As can be seen in Tables 3 and 4, all three methods perform similarly well on the relatively easy daytime queries of the RobotCar Seasons dataset. However, our approach significantly outperforms the other two methods under all other conditions. This clearly validates our idea of semantic consistency scoring.

5 Conclusion

In this paper, we have presented a method for soft outlier filtering by using the semantic content of a query image. Our method ranks the 2D-3D matches found by feature-based localization pipelines depending on how well they agree with

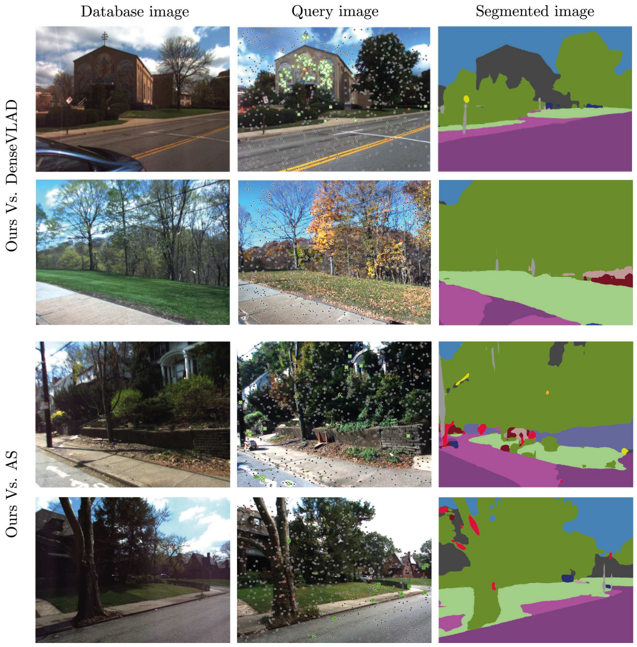

Fig. 6. Illustrations of the result of our method on the CMU Seasons dataset. Rows 1 and 3 show query images that our method successfully localizes (error $< .25$ m) while DenseVLAD and AS fail (error > 10 m) and rows 2 and 4 the vice versa. Green boxes indicate true correspondences, while gray circles indicate false correspondences. White/red crosses indicate correctly/incorrectly detected inliers, respectively.

the scene semantics. Provided that the gravity direction and camera height are (roughly) known, the camera is constrained to lie on a circle for a given match. Traversing this circle and projecting the semantically labelled scene geometry into the query image, we calculate a semantic consistency score for this match based on the fit between the projected and observed semantic labels. The scores are then used to bias sampling during RANSAC-based pose estimation.

Experiments on two challenging benchmarks for long-term visual localization show that our approach outperforms state-of-the-art methods. This validates our idea of using scene semantics to distinguish correct and wrong matches and shows the usefulness of semantic information in the context of visual localization.

Acknowledgements This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, the Swedish Research Council (grant no. 2016-04445), the Swedish Foundation for Strategic Research (Semantic Mapping and Visual Navigation for Smart Robots), and Vinnova /FFI (Perceptron, grant no. 2017-01942).

References

1. Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: CVPR (2016)
2. Arandjelović, R., Zisserman, A.: Visual Vocabulary with a Semantic Twist. In: ACCV (2014)
3. Ardeshir, S., Zamir, A.R., Torroella, A., Shah, M.: GIS-Assisted Object Detection and Geospatial Localization. In: ECCV (2014)
4. Atanasov, N., Zhu, M., Daniilidis, K., Pappas, G.J.: Localization from semantic observations via the matrix permanent. IJRR **35**(1-3), 73–99 (2016)
5. Badino, H., Huber, D., Kanade, T.: Visual Topometric Localization. In: IV (2011)
6. Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C.: DSAC - Differentiable RANSAC for Camera Localization. In: CVPR (2017)
7. Brachmann, E., Rother, C.: Learning Less is More - 6D Camera Localization via 3D Surface Regression. In: CVPR (2018)
8. Camposeco, F., Sattler, T., Cohen, A., Geiger, A., Pollefeys, M.: Toroidal Constraints for Two-Point Localization under High Outlier Ratios. In: CVPR (2017)
9. Castle, R.O., Klein, G., Murray, D.W.: Video-rate localization in multiple maps for wearable augmented reality. In: ISWC (2008)
10. Cavallari, T., Golodetz, S., Lord, N.A., Valentin, J., Di Stefano, L., Torr, P.H.S.: On-The-Fly Adaptation of Regression Forests for Online Camera Relocalisation. In: CVPR (2017)
11. Chen, D.M., Baatz, G., Köser, K., Tsai, S.S., Vedantham, R., Pylvänäinen, T., Roimela, K., Chen, X., Bach, J., Pollefeys, M., Girod, B., Grzeszczuk, R.: City-Scale Landmark Identification on Mobile Devices. In: CVPR (2011)
12. Choudhary, S., Narayanan, P.J.: Visibility Probability Structure from SfM Datasets and Applications. In: ECCV (2012)
13. Chum, O., Matas, J.: Matching with PROSAC - progressive sample consensus. In: CVPR (2005)
14. Cohen, A., Sattler, T., Pollefeys, M.: Merging the Unmatchable: Stitching Visually Disconnected SfM Models. In: ICCV (2015)
15. Cohen, A., Schönberger, J.L., Speciale, P., Sattler, T., Frahm, J., Pollefeys, M.: Indoor-Outdoor 3D Reconstruction Alignment. In: ECCV (2016)
16. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In: CVPR (2016)
17. Cummins, M., Newman, P.: Appearance-only SLAM at large scale with FAB-MAP 2.0. IJRR **30**(9), 1100–1123 (2011)
18. Fischler, M., Bolles, R.: Random Sampling Consensus: A Paradigm for Model Fitting with Application to Image Analysis and Automated Cartography. Commun. ACM **24**, 381–395 (1981)
19. Haralick, R., Lee, C.N., Ottenberg, K., Nölle, M.: Review and analysis of solutions of the three point perspective pose estimation problem. IJCV **13**(3), 331–356 (1994)
20. Irschara, A., Zach, C., Frahm, J.M., Bischof, H.: From Structure-from-Motion Point Clouds to Fast Location Recognition. In: CVPR (2009)
21. Kendall, A., Cipolla, R.: Geometric loss functions for camera pose regression with deep learning. In: CVPR (2017)
22. Kendall, A., Grimes, M., Cipolla, R.: PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In: ICCV (2015)

23. Kneip, L., Scaramuzza, D., Siegwart, R.: A Novel Parametrization of the Perspective-Three-Point Problem for a Direct Computation of Absolute Camera Position and Orientation. In: CVPR (2011)
24. Knopp, J., Sivic, J., Pajdla, T.: Avoding Confusing Features in Place Recognition. In: ECCV (2010)
25. Kobyshev, N., Riemenschneider, H., Gool, L.V.: Matching Features Correctly through Semantic Understanding. In: 3DV (2014)
26. Kukelova, Z., Bujnak, M., Pajdla, T.: Closed-form Solutions to Minimal Absolute Pose Problems with Known Vertical Direction. In: ACCV (2011)
27. Li, Y., Snavely, N., Huttenlocher, D., Fua, P.: Worldwide Pose Estimation Using 3D Point Clouds. In: ECCV (2012)
28. Li, Y., Snavely, N., Huttenlocher, D.P.: Location Recognition using Prioritized Feature Matching. In: ECCV (2010)
29. Liu, L., Li, H., Dai, Y.: Efficient Global 2D-3D Matching for Camera Localization in a Large-Scale 3D Map. In: ICCV (2017)
30. Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. IJCV **60**(2) (2004)
31. Lynen, S., Sattler, T., Bosse, M., Hesch, J., Pollefeys, M., Siegwart, R.: Get Out of My Lab: Large-scale, Real-Time Visual-Inertial Localization. In: RSS (2015)
32. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 Year, 1000km: The Oxford RobotCar Dataset. IJRR **36**(1), 3–15 (2017)
33. Massiceti, D., Krull, A., Brachmann, E., Rother, C., Torr, P.H.: Random Forests versus Neural Networks - What's Best for Camera Relocalization? In: ICRA (2017)
34. Neuhold, G., Ollmann, T., Rota Bulò, S., Kontschieder, P.: The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In: ICCV (2017), https://www.mapillary.com/dataset/vistas
35. Salas-Moreno, R.F., Newcombe, R.A., Strasdat, H., Kelly, P.H.J., Davison, A.J.: SLAM++: Simultaneous Localisation and Mapping at the Level of Objects. In: CVPR (2013)
36. Sattler, T., Leibe, B., Kobbelt, L.: Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. PAMI **39**(9), 1744–1756 (2017)
37. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Kahl, F., Pajdla, T.: Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In: CVPR (2018)
38. Sattler, T., Havlena, M., Radenovic, F., Schindler, K., Pollefeys, M.: Hyperpoints and Fine Vocabularies for Large-Scale Location Recognition. In: ICCV (2015)
39. Sattler, T., Havlena, M., Schindler, K., Pollefeys, M.: Large-Scale Location Recognition and the Geometric Burstiness Problem. In: CVPR (2016)
40. Sattler, T., Torii, A., Sivic, J., Pollefeys, M., Taira, H., Okutomi, M., Pajdla, T.: Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization? In: CVPR (2017)
41. Schindler, G., Brown, M., Szeliski, R.: City-Scale Location Recognition. In: CVPR (2007)
42. Schönberger, J.L., Pollefeys, M., Geiger, A., Sattler, T.: Semantic Visual Localization. In: CVPR (2018)
43. Schönberger, J.L., Frahm, J.M.: Structure-From-Motion Revisited. In: CVPR (2016)
44. Schreiber, M., Knöppel, C., Franke, U.: LaneLoc: Lane marking based localization using highly accurate maps. In: IV (2013)

45. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. In: CVPR (2013)
46. Singh, G., Košecká, J.: Semantically Guided Geo-location and Modeling in Urban Environments. In: Large-Scale Visual Geo-Localization (2016)
47. Svärm, L., Enqvist, O., Kahl, F., Oskarsson, M.: City-Scale Localization for Cameras with Known Vertical Direction. PAMI **39**(7), 1455–1461 (2017)
48. Toft, C., Olsson, C., Kahl, F.: Long-term 3D Localization and Pose from Semantic Labellings. In: ICCV Workshops (2017)
49. Torii, A., Arandjelović, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 Place Recognition by View Synthesis. In: CVPR (2015)
50. Valentin, J., Nießner, M., Shotton, J., Fitzgibbon, A., Izadi, S., Torr, P.: Exploiting Uncertainty in Regression Forests for Accurate Camera Relocalization. In: CVPR (2015)
51. Walch, F., Hazirbas, C., Leal-Taixé, L., Sattler, T., Hilsenbeck, S., Cremers, D.: Image-Based Localization Using LSTMs for Structured Feature Correlation. In: ICCV (2017)
52. Weyand, T., Kostrikov, I., Philbin, J.: PlaNet - Photo Geolocation with Convolutional Neural Networks. In: ECCV (2016)
53. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: LIFT: Learned Invariant Feature Transform. In: ECCV (2016)
54. Yu, F., Koltun, V.: Multi-Scale Context Aggregation by Dilated Convolutions. In: ICLR (2016)
55. Yu, F., Xiao, J., Funkhouser, T.A.: Semantic alignment of LiDAR data at city scale. In: CVPR (2015)
56. Zamir, A.R., Shah, M.: Accurate Image Localization Based on Google Maps Street View. In: ECCV (2010)
57. Zamir, A.R., Shah, M.: Image Geo-Localization Based on MultipleNearest Neighbor Feature Matching Using Generalized Graphs. PAMI **36**(8), 1546–1558 (2014)
58. Zeisl, B., Sattler, T., Pollefeys, M.: Camera Pose Voting for Large-Scale Image-Based Localization. In: ICCV (2015)
59. Zhang, W., Kosecka, J.: Image based Localization in Urban Environments. In: 3DPVT (2006)
60. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid Scene Parsing Network. In: CVPR (2017)