Engineering Science and Technology, an International Journal 35 (2022) 101098

HOSTED BY

Contents lists available at ScienceDirect

Engineering Science and Technology, an International Journal

journal homepage: www.elsevier.com/locate/jestch

Full Length Article

Long-term image-based vehicle localization improved with learnt semantic descriptors

Ibrahim Cinaroglu [a,b,*,1], Yalin Bastanlar [b]

[a] Karamanoglu Mehmetbey University, Department of Computer Engineering, Karaman 70100, Turkey
[b] Izmir Institute of Technology, Department of Computer Engineering, Urla, Izmir 35433, Turkey

ARTICLE INFO

Article history:
Received 20 June 2021
Revised 3 December 2021
Accepted 19 January 2022
Available online xxxx

Keywords:
Image-based localization
Image matching
Autonomous driving
Semantic segmentation
Semantic descriptor

ABSTRACT

Vision based solutions for the localization of vehicles have become popular recently. In this study, we employ an image retrieval based visual localization approach, in which database images are kept with GPS coordinates and the location of the retrieved database image serves as the position estimate of the query image in a city scale driving scenario. Regarding this approach, most existing studies only use descriptors extracted from RGB images and do not exploit semantic content. We show that localization can be improved via descriptors extracted from semantically segmented images, especially when the environment is subjected to severe illumination, seasonal or other long-term changes. We worked on two separate visual localization datasets, one of which (Malaga Streetview Challenge) has been generated by us and made publicly available. Following the extraction of semantic labels in images, we trained a CNN model for localization in a weakly-supervised fashion with triplet ranking loss. The optimized semantic descriptor can be used on its own for localization or preferably it can be used together with a state-of-the-art RGB image based descriptor in hybrid fashion to improve accuracy. Our experiments reveal that the proposed hybrid method is able to increase the localization performance of the standard (RGB image based) approach up to 7.7% regarding Top-1 Recall values.

1. Introduction

Visual localization (VL) can be defined as estimating the position and orientation of a visual query material within a known environment. Information on location of a mobile device (could be a pedestrian or a vehicle) is critical for city-scale navigation and other location-based services. Also due to the limitations of GPS-based localization in urban environment (e.g. signal failure in a cluttered environment), visual localization attracted an increasing attention in the last decade [33].

In our work, an image retrieval based VL technique (Fig. 1) is employed with an approximate nearest neighbor search algorithm. This method utilizes a database of geotagged images and the known geographic location of the retrieved database image (best match) serves as the position estimate of the query image.

The approach proposed in this paper is based on the hypothesis that semantic decomposition of a scene can increase localization performance. We can rely on the semantic labels especially when there are long-term changes in the scene. As can be observed in Fig. 2, with illumination conditions (sunny, cloudy etc.) and seasonal variances (summer, winter etc.) drastic appearance changes occur. Standard appearance based methods face difficulties in such cases, whereas semantic segmentation can give stable results. Therefore, using this superior ability of semantic knowledge to understand a scene has been our main motivation in this study.

Some related studies [53,40,29,25,43] showed that semantic cues can be used to improve localization accuracy but none of them have directly performed localization using a descriptor extracted from semantically segmented images. Novelty in our study is that we improve the localization performance directly using learnt semantic descriptors trained with semantically segmented images. More specifically, we take a state-of-the-art appearance-based localization method where local descriptors are extracted from RGB images (LD-VL) and combine it with our newly developed semantic-descriptor based method (SD-VL), resulting in a novel hybrid localization approach. We have worked on datasets where the environment is subjected to illumination and other long-term changes and observed a performance improvement with the proposed hybrid approach.

We summarize the main contributions of our work as follows:

* Corresponding author.
E-mail addresses: ibrahimcinaroglu@kmu.edu.tr (I. Cinaroglu), yalinbastanlar@iyte.edu.tr (Y. Bastanlar).
[1] This author has just changed institution and currently works at Karamanoglu Mehmetbey University.

https://doi.org/10.1016/j.jestch.2022.101098
2215-0986/

Fig. 1. On the left, we see a query image. On the right, we see a district with a database of images with known GPS coordinates. Retrieval from the database is based on the similarity of descriptor vectors. The GPS location of the image retrieved from the database serves as the position estimate of the query image. If the estimate is within a certain distance limit, then the localization is considered as successful.

- A novel semantic descriptor is trained with a CNN model that includes NetVLAD [3] layer, using semantically segmented images as input. Then, this optimized semantic representation is used directly for visual localization (SD-VL).
- We generated the *Malaga Streetview Challenge* dataset based on the Google Streetview, which provides wide baseline and severe environmental changes together. This newly generated test set has been made publicly available and we believe it will be useful for researchers studying in this field.
- Hybrid-VL is proposed to combine newly developed SD-VL and the baseline LD-VL methods in post-processing stage. We experimentally show that the proposed hybrid method increases localization performance measured with frequently used evaluation metrics *Top-1 Recall@D* and *Recall@N* on *Malaga Streetview Challenge* and *RobotCar Seasons* (a benchmark dataset for visual localization).

The remainder of this paper is structured as follows. The related works are reviewed in Section 2. Section 3 provides detailed information about our method. Both experimental results and details of dataset preparation can be found in Section 4 which is followed by the conclusions in Section 5.

2. Related work

2.1. Localization with appearance based descriptors

Classical methods of image retrieval based localization mostly depend on Bag-of-Features [32,44] approach. Furthermore, this approach generally has expressed an image with local descriptors that are created from points of interest. Scale-Invariant Feature Transforms (SIFT [23]) can be given as a frequently used example of these local descriptors. In this local descriptor, distinctive invariant features were extracted from images which also can be used to carry out reliable matching between changing views of an object or scene.

In this Bag-of-Features approach, SIFT-like descriptors extracted from all images in the database are clustered to define a set of 'visual words', then an image is represented by a visual word frequency histogram. The similarity between two images is measured by the distance between their histogram vectors. In time, researchers managed to perform the same task with less memory [19] and gain robustness to repetitive structures [50], illumination, viewpoint changes and long-term changes [49]. This approach has been implemented for 360 degree panoramic images as well [17,30].

Recent studies consider using features from the deep convolutional layers of convolutional neural networks (CNNs) [46,9]. A trainable CNN, NetVLAD, was proposed by Arandjelovic et al. [3] in which a specially designed layer is added to a standard CNN to convert the last convolutional layer into a compact descriptor. In their study, NetVLAD outperformed state-of-the-art localization

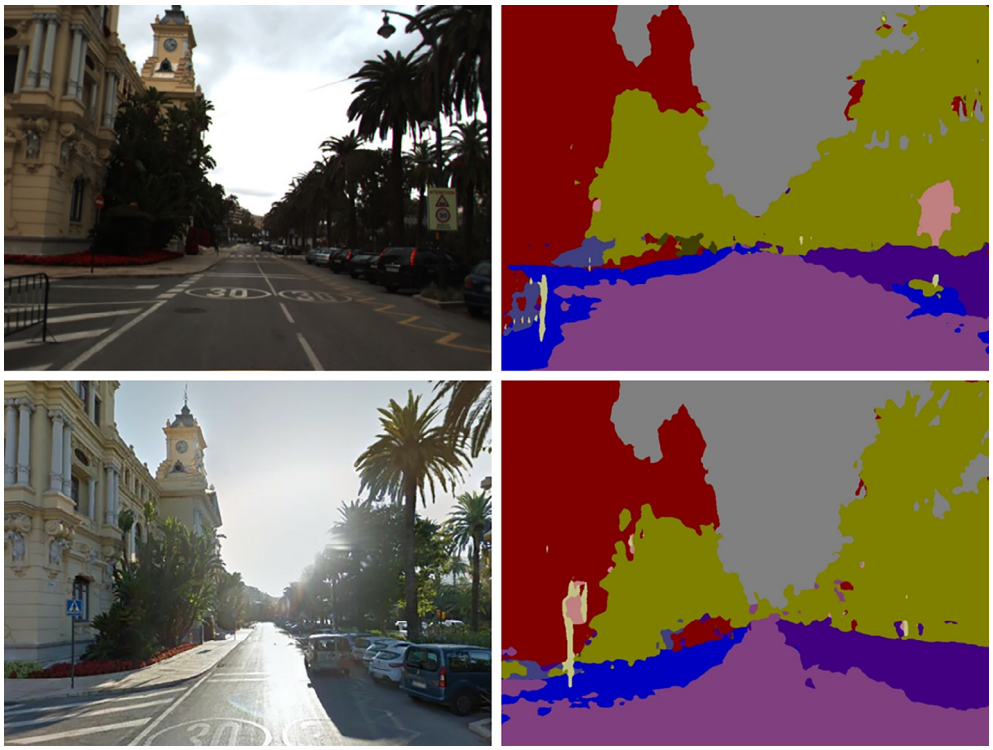

Fig. 2. On the left, two images of the same scene with considerable illumination changes. On the right, their semantic segmentation results. Standard methods have low performance for such cases, where more stable semantic segmentation can help.

techniques based on experiments performed on four different datasets.

There are other powerful descriptors for image retrieval such as Regional-MAC [48], Generalized Mean Pooling [36] and Local Texton XOR Patterns [4]. However, these were developed especially for place recognition (e.g. Is this the Eiffel Tower?) rather than localization. When localization (up to a distance threshold) is considered, NetVLAD is still one of the best methods [34]. Therefore, in our study we have employed NetVLAD as the baseline approach.

Studies given above extract features from RGB images only. We refer to them as RGB image based or appearance based methods.

2.2. Using semantic labels for long-term localization

CNN based semantic segmentation approaches have achieved impressive results in different computer vision tasks by using both standard and larger field-of-view cameras [20,16,7,31]. Also the idea of using semantic labels to improve image based localization has been explored before. In [25], localization is based on standard feature point descriptors but feature points not belonging to man-made objects (e.g. trees) are considered as unreliable and they are eliminated via semantic information. In [29], features are extracted from the convolution layers of a CNN, but a weighting scheme is applied based on semantic labels (e.g. increasing weights for buildings since they are more stable in long term). Seymour et al. [40] developed a deep learning based method for fusing appearance and semantic information. They proposed an attention module to predict the most reliable regions of appearance and semantic modalities.

An attempt to design a descriptor from 2D semantic labels was first proposed in [43] but rather than localization, the descriptor was used to distinguish street intersections from other scenes. Also a framework was proposed in [53] that uses semantic edge features from images to achieve on-road localization. Firstly in our previous work [11], we optimized a semantic descriptor based on the semantic labels of the entire image and performed localization with that descriptor rather than using it as a clue. In this paper, we extend our previous work by combining it with a state-of-the-art appearance based method (NetVLAD) and exceed its performance.

Some previous work on image-based VL fall into the category of 3D structure-based localization which employs a 3D model of the scene to match with the information extracted from images. Stenborg et al. [45] performed localization based on the query image's semantic content when the environment is 3D reconstructed and semantically labeled. This is an innovative study in terms of performing localization purely based on semantic labels; however it requires semantic labels of 3D point clouds, which is not available in most cases. In another example, 2D-3D point matches are checked if their semantic labels are also matching [47]. In [38], a dictionary is developed for semantic content and the scene is represented as Bag-of-Semantic-Words. Our method is based on 2D images and their semantic segmentation results. It is much cheaper than the localization approaches that require the semantic 3D reconstruction of the environment. In addition, it was reported in several previous studies [51,6] that 2D approaches perform as well as 2D-3D matching approaches.

2.3. Other modalities for long-term localization

Some previous work exploited modalities other than semantic labels to handle illumination and long-term variations. Piasco et al. [34] used geometry information while training their new global image descriptor. They managed to increase localization performance thanks to the depth map belongs to each query image. Germain et al. [15] also produced a global image descriptor by adding condition-specific sub-networks to a state-of-the-art CNN based image retrieval architecture. Their descriptor is computed according to capturing condition and becomes successful especially against day-night variation. Again in order to cope with night-to-day challenge, Anoosheh *et al.* [2] managed to increase localization accuracy via converting nighttime driving images to a daytime representation thanks to their novel image translation model ToDayGAN. Also, Porav *et al.* [35] proposed an invertable generator that is able to convert the conditions of images to a desired opposite ones. Their trained network outputs synthetic images to manage this appearance transferring which is designed to help standard local feature matching method SURF. Doan *et al.* [13] introduced a new Monte Carlo localization algorithm based on image retrieval. Moreover they proposed a software that works with role playing game in order to collect hyper-realistic computer-generated imagery of a city from the street level for providing different environmental conditions.

3. Our method

3.1. Method overview

The proposed semantic VL approach is also based on the image retrieval technique previously depicted in Fig. 1. However, we introduce semantic descriptors to find the best match instead of standard appearance based descriptors. Thus, database consists of pixel-level semantic segmentations of geotagged images (Fig. 3). Similar to most visual localization studies in the literature, our prior map (yellow path) corresponds to reference traversal of the dataset, while images of other traversals on the same path but collected in changing conditions are our query images.

To learn the best possible semantic descriptor, we train a CNN model using a section of the route which is dedicated to training. Test results are obtained with an unseen section, i.e. training and test samples are geographically disjoint.

The proposed Hybrid-VL method can be summarized step by step with the given pseudo-code in Fig. 4 which is constructed on image retrieval in 2D-2D matching space. This reductionist representation of proposed VL method also provides us which step corresponds to which key components of a characteristic image retrieval based localization system (image representation, image matching, hybridization). In addition, we are able to show not only where the novel parts of this study takes place with their corresponding steps, but also how (offline-online) these parts are operated.

In this paragraph, the proposed algorithm to match input images with geotagged ones is introduced. Firstly note that the algorithm from line 1 to 8 can and should be computed offline, regarding to an actual driving mission. In this representation, proposed learnt SD-VL method takes a query image I_a and return k number of candidates C_a^{SD} from database images in lines from 1 to 13. In the first line previously retrained semantic segmentation method *DeepLabv3 + Retrained* is employed on database images *I*that gives us their segmented versions *S*. In line 2, a CNN model is trained on *S*with triplet ranking loss for VL task, then the part from line 3 to 6 corresponds to learnt semantic descriptor SD_i extraction process. Robust indexing is built up in line 7 with our ANNS method FLANN for database image descriptors collection SD_T. Next, these same steps are operated for a semantically segmented query image S_a in line 9 and 10. From line 11 to 13, ANNS is conducted and *k*number of best matching images retrieved. Moreover, the same steps (2–12) for SD-VL method is repeated without segmentation in order to obtain learnt LD-VL method in lines 14 and 15, so that we obtain best matching *k*candidates C_a^{LD} for our LD-VL method. Finally, effective decision-level hybridization methods is represented from line 16 to 18, that incorporates

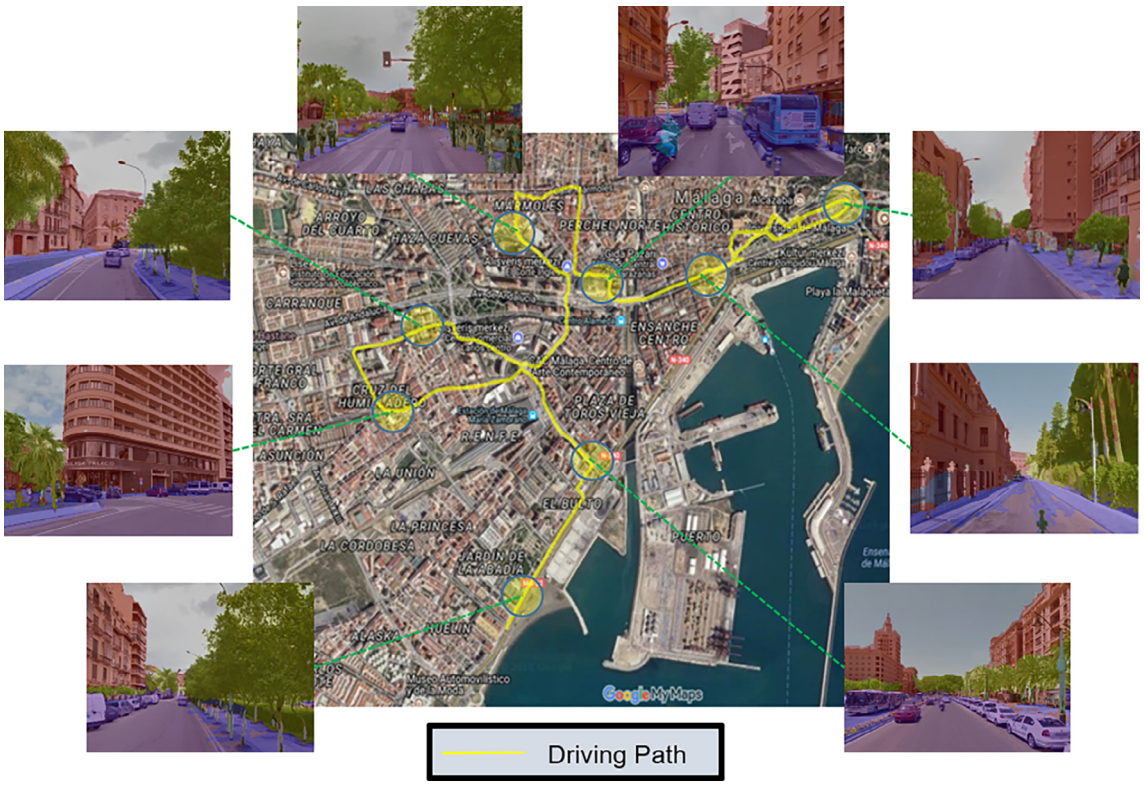

Fig. 3. In the proposed semantic content based VL approach, the database consists of pixel-wise semantic segmentations of geotagged images.

C_a^{LD} and C_a^{SD} in post-processing level. As a result, among the *k*number of Hybrid-VL method candidates C_a^{hybrid} the top first one I_1^{hybrid} is returned against the given query image I_a.

In the following sections, we give further explanation about the proposed algorithm with detailed implementation of each step.

3.2. Semantic segmentation of geotagged images

Pixel-level semantic labels of database and query images are extracted with the state-of-the-art DeepLabv3+ [8] model which was pretrained on CamVid dataset [14]. In order to increase the performance of pretrained DeepLabv3+, we conducted a retraining with images in our dataset. While retraining, the absence of annotated ground truth images for the *RobotCar Seasons* dataset encouraged us to follow a weakly-supervised approach. To be more specific, successful segmentation results of *Pretrained DeepLabv3 +* on *RobotCar Seasons* dataset are accepted as annotations and they are used to retrain the model. In this way, we generated adequate amount of labeled images in the target dataset without requiring manual annotation. Steps of this weakly-supervised retraining are given below:

- All the images in the query sets of *RobotCar Seasons* dataset (cf. Table 1) except for *night* and *night-rain* sets were semantically segmented with the *Pretrained DeepLabv3 +* model. This totals to 2500 + images.
- Segmentation results that reflect our semantic classes in the best way were manually selected. About 170 images per query set were selected in this way with a total of 1024 labeled images. These selected images were excluded from localization experiments (query sets) since they have been seen by our retrained segmentation CNN.
- *DeepLabv3 +* was retrained with these 1024 images and this new model is named as *DeepLabv3 + Retrained*.

Retrained model produced satisfactory segmentation performance, examples of which can be seen in Fig. 5. Our model classifies the pixels into 11 semantic classes (Building, Car, Road, Sidewalk, Sky, Tree, Pedestrian, Bicycle, Pole, Fence, Sign Symbol) as depicted in the figure.

3.3. Training semantic descriptors for localization

In the past, we had implemented the idea of designing semantic descriptors manually, where we divided the images into 4 equal pieces and put the class frequencies in those pieces into vectors [10]. The results were not very satisfactory and it is obvious that the ideal solution is learning a semantic descriptor automatically using a dataset contains the targeted long-term variations and illumination changes.

With this aim, our semantically segmented database images are given to a CNN as the training set to minimize a triplet loss function (Fig. 6). In triplet loss, firstly introduced in FaceNet [39], for a given input image (anchor), images taken from a similar location constitute a positive set and images from far away positions constitute a negative set as visualized in Fig. 7. By training with triplet loss, a descriptor (last layer of CNN) is optimized so that the distance to the positive set is minimized and the distance to the negative set is maximized. We use AlexNet [22] as our backbone CNN with addition of NetVLAD layer in order to obtain our learnt semantic descriptor. Actually we also examined the VGG16 [42] as a deeper and up-to-date network, yet we prefer to use AlexNet due to its better localization performance. In fact, it is an expected result that a less complex CNN such as AlexNet gives better results in semantic descriptor based localization, in which the features are extracted from a simple representation (semantic labels). The employed triplet ranking loss will be explained next.

Desired location-aware descriptor is represented with $f_\theta(q) \in \mathbb{R}^d$ where a query image *q*is embedded into a *d*-

Input: A finite set $I = \{I_1, I_2, \ldots, I_n\}$ of ground geotagged database images
Input: A query image I_a taken while driving in a street-like environment
Output: The location of the vehicle in the prior map and the best match I_1^{hybrid}, respectively

Computed Offline	SD-VL	1	S = **segment** I semantically using CNN based '*DeepLabv3+ Retrained*' model;
		2	$f_S^{triplet}$ = **train** CNN model on S_{train} and S_{val} with triplet ranking loss for VL task;
		3	SD_T = *learnt* descriptor collection of all images in S;
		4	**for** $i \leftarrow 1$ **to** n **do**
		5	SD_i = extract image features using $f_S^{triplet}(S_i)$;
		6	**add** SD_i **to** SD_T
		7	**build** index on SD_T using FLANN ;
		8	k = 10 number of retrieved candidate;
Computed Online		9	S_a = semantically segmentation of I_a using CNN based '*DeepLabv3+ Retrained*' model;
		10	SD_a= extract image features using $f_S^{triplet}(S_a)$;
		11	**search** approximate nearest-neighbor feature matches for SD_a in SD_T: $C_a = ANNS\ (SD_a, SD_T)$;
		12	**select** k first image matches $I^p \subseteq C_a$: $I^p = \{I_1^p, I_2^p, \ldots, I_k^p\}$;
		13	$C_a^{SD} \leftarrow I^p$: nearest candidates for I_a using SD_a;
	LD-VL	14	**repeat** the line from 2-12 on I_a (RGB) without segmentation with $f_{RGB}^{triplet}$, LD_T and LD_a;
		15	$C_a^{LD} \leftarrow I^p$: nearest candidates for I and I_a using LD_a;
	Hybridization	16	C_a^{hybrid} = **hybrid** (C_a^{SD}, C_a^{LD}) ;
		17	$I_1^{hybrid} \leftarrow$ **select** first candidate in C_a^{hybrid};
		18	**return** I_1^{hybrid};

Fig. 4. Proposed algorithm of Decision-level Hybrid-VL.

Table 1
Detailed statistics for the two benchmark datasets used in this study.

Dataset	Baseline	Database images conditions (# images)	Query images conditions (# images)
RobotCar Seasons [37]	Short baseline	Overcast-Reference (6954)	dawn (483), dusk (394), night (483), night + rain (440), rain (421), overcast summer (463), overcast winter (390), snow (489), sun (460)
Malaga Streetview Challenge (ours)	Wide baseline	Reference traversal in 2014 (Overcast/ 1561)	Google Streetview (436): all short-long term changes by different time period and years from 2014 to 2020

dimensional Euclidean space. Here, θ corresponds to training parameters that are to be optimized. In this aim, from the dataset we use (*RobotCar Seasons* or *Malaga Streetview Challenge*), we acquire a training set of tuples $(q, p^q, \{n_j^q\})$, where for each training query image q we have a positive p^q (closest image) and a set of definite negatives $\{n_j^q\}$ (metric distance to the query is higher than a threshold). We select p^q as the closest image in the database t^{db} according to the GPS coordinates:

$$p^q = \underset{t^{db}}{\operatorname{argmin}}\ d_{gps}(q, t^{db}), \tag{1}$$

which is slightly different from the original NetVLAD implementation [3], where closest image is selected from a set of potential positives. That was because they used Google Streetview images looking at different directions and did not know which correctly located image actually had a view overlapping with the query image.

Let $d_\theta(q, p^q) = ||f_\theta(q) - f_\theta(p^q)||$, then the objective becomes to learn the training parameters θ so that distance between the query q and the positive image p^q is smaller than the distance between the query q and all negative images in $\{n_j^q\}$:

$$d_\theta(q, p^q) \ < \ d_\theta(q, n_j^q), \quad \forall j. \tag{2}$$

Finally, triplet ranking loss L_θ is defined as

$$L_\theta = \sum_j h\Big(d_\theta^2(q, p^q) + m - d_\theta^2(q, n_j^q)\Big), \tag{3}$$

where h is the hinge loss $h(x) = max(x, 0)$, and m is a margin that determines the amount of dissimilarity between positive and negative pairs (Fig. 7). According to Eq. (3), if the squared distance of a negative image is greater than (by a margin) the squared distance of the positive image, the loss is zero. Otherwise, loss increases proportional to the amount of violation. In this way, our '*learnt*' descriptor becomes end-to-end trainable. The above described pro-

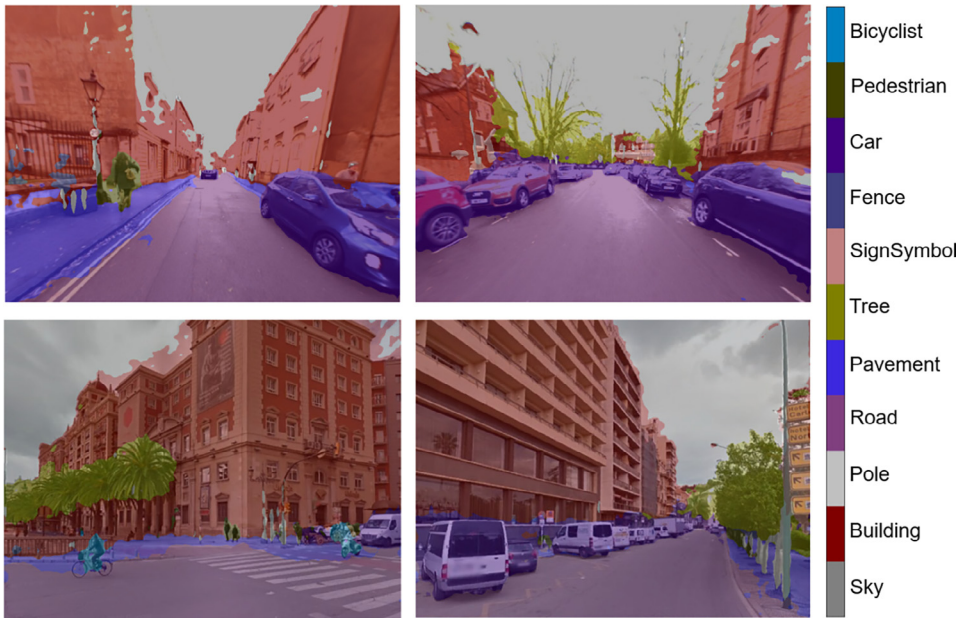

Fig. 5. Semantic segmentation performance of the *DeepLabv3 + Retrained* on some sample images of *RobotCar Seasons* (1st row) and *Malaga Streetview Challenge* (2nd row) dataset.

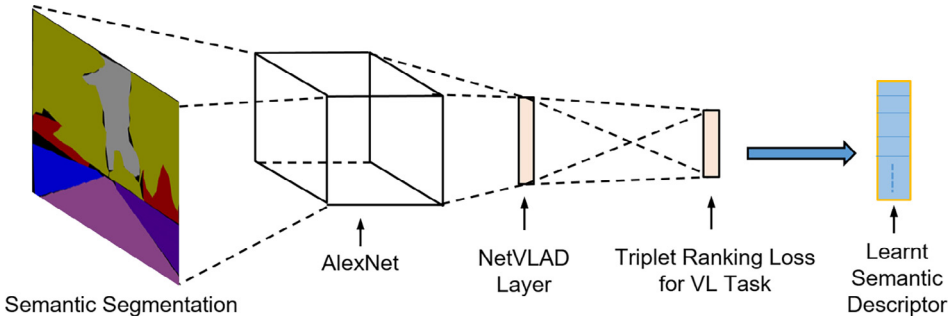

Fig. 6. *Learnt* SD (16 k) trained with triplet ranking loss for VL task on semantically segmented images.

Fig. 7. An anchor image together with positive (same location) and negative (different location) samples to be used in triplet loss while training localization CNN.

cedure is considered as a weakly-supervised training since no human supervision is employed, the annotations can be noisy due to position shifts and overlap between views can be limited.

The described triplet ranking loss based learning process gives us a learnt semantic descriptor when semantic labels are given as input which in turn used for SD-VL.

3.4. Training appearance based descriptors for localization

We should note that, same training procedure in previous section is followed directly on RGB images in implementation of the LD-VL method. Again we examined both the VGG16 and AlexNet as a backbone for obtaining our appearance based learnt descriptor. Contrary

to the semantic descriptor case, VGG16 gives better localization performance. It is not surprising that a more complex CNN such as VGG16 gives better results in the LD-VL method, in which relatively complex information (RGB image) is used as an input.

3.5. Hybridization

As a main contribution of this study, a novel Hybrid-VL method is proposed by combining SD-VL and LD-VL methods with the aim of alleviating the drawbacks of both methods.

3.5.1. Descriptor matching

Before explaining our hybridization methodology, we first need to introduce our descriptor matching approach. We use the previously learnt parameters (Section 3.3) to extract the representations of all database images $\{I\}$ which can be done offline and denoted by $f_\theta(I)$. In test time, we need to perform an efficient comparison between the collection $f_\theta(I)$ and a given query $f_\theta(q)$ to find the nearest database image. This comparison task emerges as one of the most important steps of a typical VL system. To deal with this task efficiently, different types of fast approximate nearest neighbor search (ANNS) methods [12,52,18] were introduced, some of which are CNN-based [1].

In short, ANNS methods look for the approximate nearest neighbors instead of exact nearest neighbors while comparing the elements of large databases in computer vision applications. Furthermore, superiority of using k-dimensional (k-d) trees in ANNS was highlighted in previous works [41,21]. Additionally, Muja and Lowe [27,28] improved these k-d trees by randomizing them which is named as *multiple randomized k-d trees*. Also, they mapped their efficient method into a compact tool called fast library for ANNS (FLANN, [26]). In our study, both LD-VL and SD-VL methods are constructed on FLANN to retrieve the most similar database image for a given query.

First of all, FLANN builds a powerful index on our database descriptors collection by means of multiple randomized k-d trees. Then ANNS is applied for each of given element in query descriptors collection by using the previously created index. Finally, it returns k nearest candidate images with their Euclidean distances (L2-norm) to the corresponding query image. In our work, k is set to 10. Let SD_i and LD_i represent the lists of nearest candidate images obtained by the two methods for a given i^{th} query image. Then, $D_j(SD_i)$ and $D_j(LD_i)$ are the corresponding distance vectors where j refers to the database index of the nearest candidate images. This collection of distance values are used to generate the final list for Hybrid-VL as described in the following section.

3.5.2. Decision-level hybrid-VL

After k number of matching results are obtained via ANNS for both SD-VL and LD-VL approaches, they are combined according to the ranks of the candidate images and their distance values.

First, in order to achieve a reliable hybridization, we normalize distance values into $[0-1]$ range, then apply histogram equalization. After this pre-processing stage, we combine SD_i and LD_i results as shown in Fig. 8. More specifically, we integrate distance values ($D_j(SD_i)$ and $D_j(LD_i)$) per each query which were previously returned in ascending order. While integrating, distance values are weighted with their own rank (higher rank candidate is penalized less) and multiplied by W or $(1-W)$, where W represents the weight of candidates coming from SD-based approach. This hybrid distance updating equation is given below:

$$D_j(i) = \begin{cases} D_j(SD_i)\cdot(\frac{rnk_j(SD_i)}{k})\cdot W + D_j(LD_i)\cdot(\frac{rnk_j(LD_i)}{k})\cdot(1-W), & \text{if } j \in (SD_i \cap LD_i) \\ D_j(LD_i), & \text{else if } j \in (LD_i) \\ D_j(SD_i), & \text{else if } j \in (SD_i) \end{cases} \quad (4)$$

where $rnk_j(SD_i)$ and $rnk_j(LD_i)$ denote the ranking of the candidate image j in SD_i and LD_i lists. In the first case, a database image is in the 10 nearest neighbor list of both methods, $j \in (SD_i \cap LD_i)$. Here W parameter enables us to tune the contribution of SD-VL and LD-VL. For instance, with $W < 0.5$, we trust more on LD-VL method.

Multiplying distances directly with their rankings $\frac{rnk_i}{k}$ (for higher rank candidates distance values are decreased more), can also be seen as rewarding the case that a candidate is found in both lists.

As a result of this update process, we obtain a combined list of $D_j(i)$ for the i^{th} query. Finally, we reorder these and accept the top 10 images in this new list as the final result of Hybrid-VL method.

4. Experiments

4.1. Datasets

We performed experiments on two datasets, both contain illumination changes and other long-term appearance changes (such as sunny/cloudy weather or occurrence of new structures). One of them is publicly available[2] and commonly used *RobotCar Seasons* dataset on which recent examples of effective VL studies [15,40,34,2,35,6] evaluated their performance. The other one, *Malaga Streetview Challenge* dataset, was prepared by us in order to test the performance of our method not only on short/long term changes but also on wide baseline as depicted in Fig. 9. We have made this dataset publicly available[3] with its geotags.

RobotCar Seasons Dataset [37] is a subset of RobotCar dataset [24] which had been collected in Oxford, UK by passing through the same 10 km route more than 100 times in a year. *RobotCar Seasons* dataset provides less variability in viewpoints (baseline) but a larger variance in viewing conditions for a city-scale urban driving scenario as summarized in Table 1. A triple camera (left, right, rear) setup was used originally. In this study we used only rear images since the driving direction is in the center of the image. In this way 6954 database images (overcast reference) were obtained. For the query set we used overcast-winter set consisting of 390 images since these provide enough amount of seasonal and illumination changes.

Malaga Streetview Challenge Dataset contains a reduced subset of publicly available Malaga Downtown Dataset [5] as the database images. These were collected on nearly 8 km. urban route visualized in Fig. 3. To be able to include viewpoint variety and long term changes, we collected query images from Google Streetview within every 10–20 meter in the same 8 km. route in different times (left column in Fig. 9). In total, *Malaga Streetview Challenge* has 436 query images and 1561 database images (Table 1).

4.2. Evaluation metrics

In this study, GPS based metric error is computed to evaluate the performance of SD-VL, LD-VL and Hybrid-VL methods. Each database and query image is associated with a GPS position, which is in WGS84 geographic coordinate system. Just summing up or averaging the metric error values to measure the localization performance is not reliable, because similar descriptor mismatching cases may result in very different GPS based metric errors. Thus, more reliable evaluation metrics were proposed for localization tasks and have been frequently used in the literature [3,49,51,34,33,15,38,37,53]. These evaluation metrics are explained below:

[2] https://data.ciirc.cvut.cz/public/projects/2020VisualLocalization/RobotCar-Seasons/

[3] https://github.com/ibrahimcinaroglu/Malaga-Streetview-Challenge

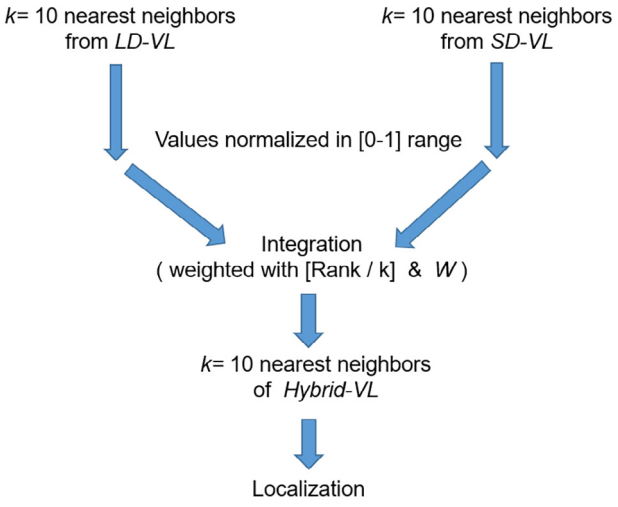

Fig. 8. Decision-level hybridization methodology.

- ***Top-1 Recall @D*:** Distance between the top ranked (1^{st}) returned database image position and the query ground truth position is calculated. Then the percentage of queries with distances less than a fix threshold D (changing from 5 to 150 meter) is plotted.
- ***Recall @N*:** Percentage of well localized ($\leqslant$25 m distance error) queries are plotted with respect to N number of returned candidates. Even if one of these N candidates is well localized then the query is accepted as correctly localized.

4.3. Experimental results

Adapting the triplet ranking loss for our descriptors requires to divide our driving path into three geographically disjoint parts as training, validation and testing set. For instance, each division of *RobotCar Seasons* contains around 2300 database and 130 query images. For the sake of fairness, all the examined VL methods were examined on the same division of *Overcast-Winter* traversal in *RobotCar Seasons* and *Malaga Streetview Challenge*.

As a result of training, we obtained 16 k dimensional VLAD vectors (Fig. 6) with $K = 64$ cluster numbers [3]. This descriptor size was used for both SD-VL and LD-VL methods. Eventually, we examined our proposed Hybrid-VL method on the test sets of *RobotCar Seasons* (130 test queries) and *Malaga Streetview Challenge* (111 test queries).

If we examine the proposed method in terms of its efficiency, steps from line 1 to 8 in the algorithm (Fig. 4) are computed offline, regarding to our actual driving mission. Then, returning a best matching database image I_1^{hybrid} for a given query image I_a (steps from line 9 to 18) takes nearly 1.5 s. Most of this time is spent between line 9 and 10 while computing the returned candidates for both LD-VL and SD-VL method. Negligible time is spent on the remaining steps (11–18).

Fig. 10 depicts the superiority of the proposed Hybrid-VL method via previously given evaluation metrics (*Top-1 Recall@D*, *Recall@N*). Hybrid-VL is able to increase *Recall@1* of LD-VL method by 4% and 3.6% on *RobotCar Seasons* (bottom-left plot) and *Malaga Streetview Challenge* (bottom-right plot) respectively. In these plots, distance threshold D is set to 25 m. which is common in related studies. As N increases, recall values increase for all meth-

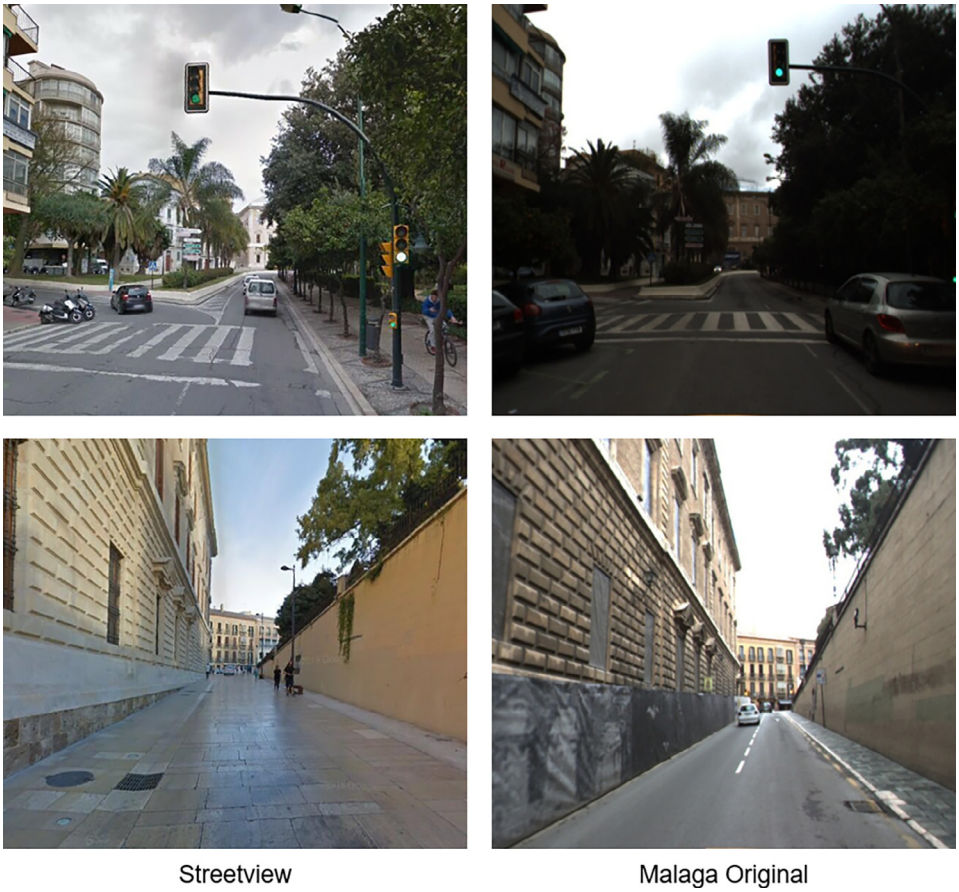

Fig. 9. Viewpoint, illumination and other long-term (new buildings, road etc.) changes between *Malaga Streetview Challenge* query images (left) and corresponding *Malaga Downtown* database images (right).

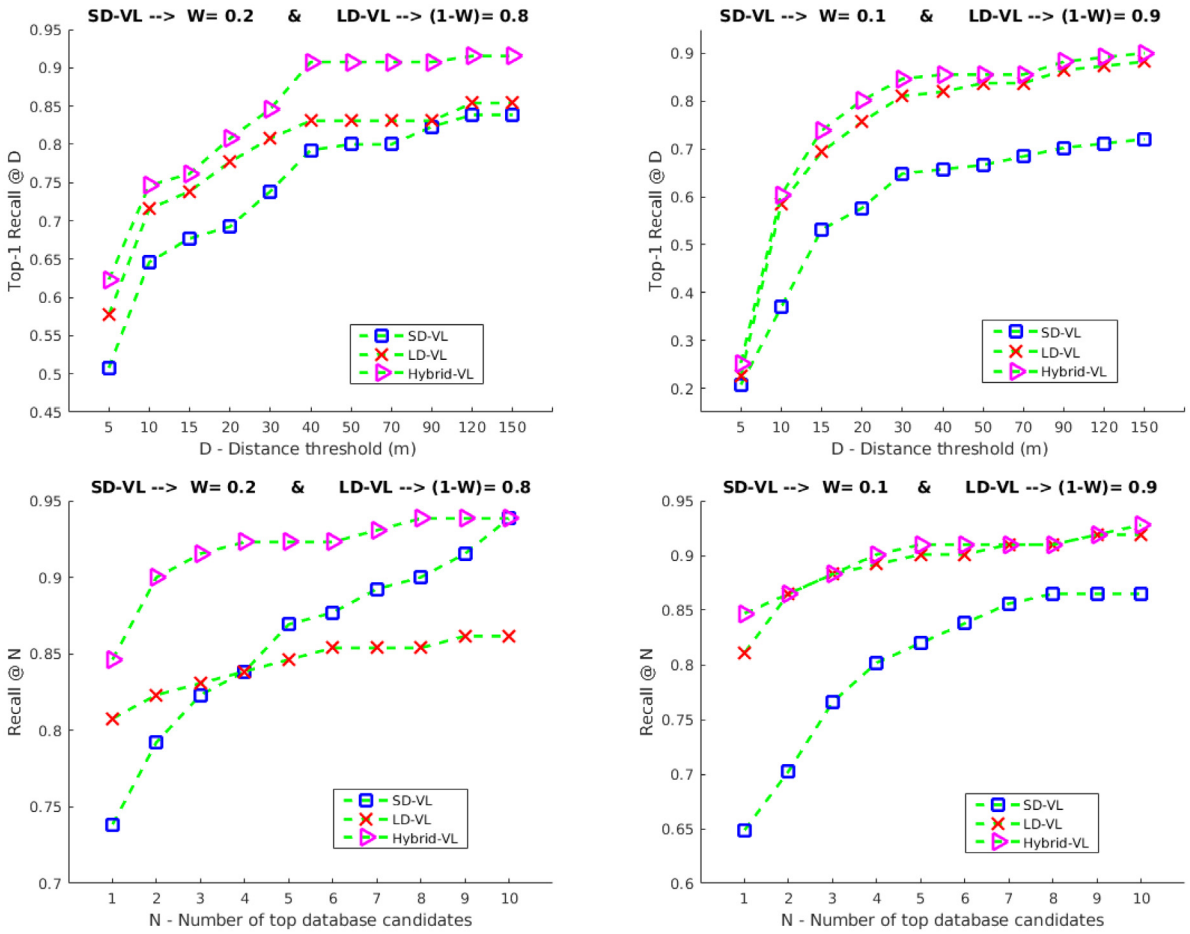

Fig. 10. Superiority of proposed Hybrid-VL method that incorporates LD-VL and SD-VL methods. Results represented with *Top-1 Recall@D* (1^{st} row) and *Recall@N* (2^{nd} row) evaluation metrics on the *RobotCar Seasons Overcast-Winter* traversal (1^{st} column) and *Malaga Streetview Challenge* (2^{nd} column).

ods but Hybrid-VL stays on the top. We are also able to observe *Top-1 Recall* values for varying distance thresholds in Fig. 10. The improvement of the proposed hybrid method over LD-VL for *RobotCar Seasons* (top-left plot) is clear. For small D values, increase in recall is around 5%. Whereas, for larger values ($D > 30$m.) the increase is more significant and reaches up to %7.7. For *Malaga Streetview Challenge* (top-right plot) the improvement over LD-VL is relatively small but still it increases the performance for every D value.

Some visual examples where LD-VL fails but Hybrid-VL approach retrieves correct locations can be viewed in Fig. 11. One can observe the challenging illumination conditions or appearance differences (changing cars). Steady semantic content helps the proposed hybrid method for better retrieval performance.

To sum up, experimental results indicate that the performance of the proposed Hybrid-VL method is superior against the state-of-the-art baseline LD-VL method (NetVLAD with RGB images) on the examined data sets. Thus, our initial hypothesis described in Section 1 'semantic decomposition of a scene can increase localization performance' is validated with the provided performance scores and visual samples.

These results were obtained with selected W parameter (0.2 for *RobotCar Seasons*, 0.1 for *Malaga Streetview Challenge*). The following subsection investigates the sensitivity of success on W parameter.

4.4. Sensitivity on W parameter

We already explained (Section 3.5.2) that the proposed Hybrid-VL method is based on W parameter. Thanks to W parameter (Eq. (4)), we are able to tune the contributions of SD-VL and LD-VL in hybridization. Logically, we should trust on the better-performing VL method among the SD-VL and LD-VL methods and LD-VL results are better in most cases. This intuition is approved in Fig. 10, where the best Hybrid-VL results are obtained by increased contribution of LD-VL ($W < 0.5$).

Results for varying values of W are given in Fig. 12. Only the results on the *RobotCar Seasons* were included in the figure but the same trend occurs for *Malaga Streetview Challenge* as well. One can observe that W values close to 0.2 (e.g. 0.1 or 0.3) result in a similar performance. In fact, as long as we trust LD-VL more than SD-VL it is beneficial to Hybrid-VL. In contrast, as we give

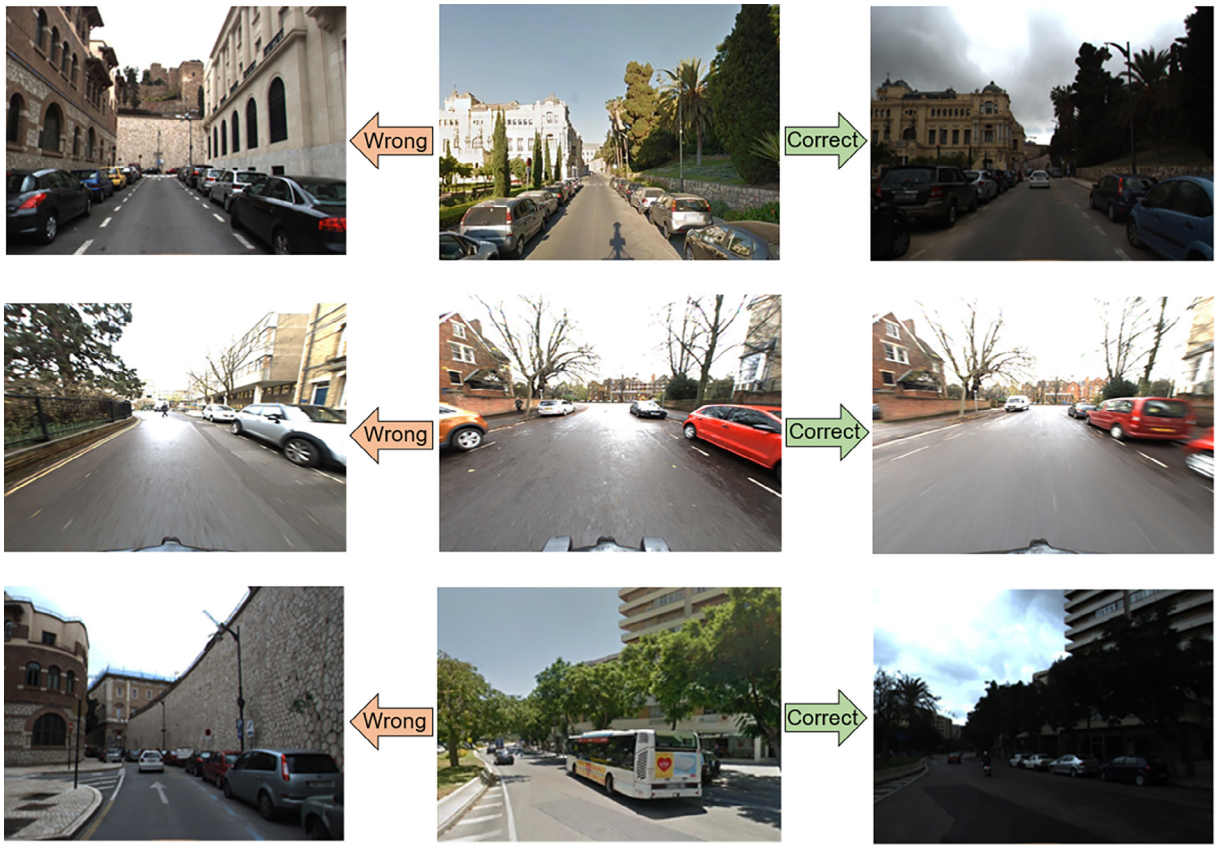

Fig. 11. Superiority of proposed Hybrid-VL method with three sample localization cases from both datasets. RGB image based method LD-VL (left) fails but Hybrid-VL (right) retrieves correct images for a given query (middle).

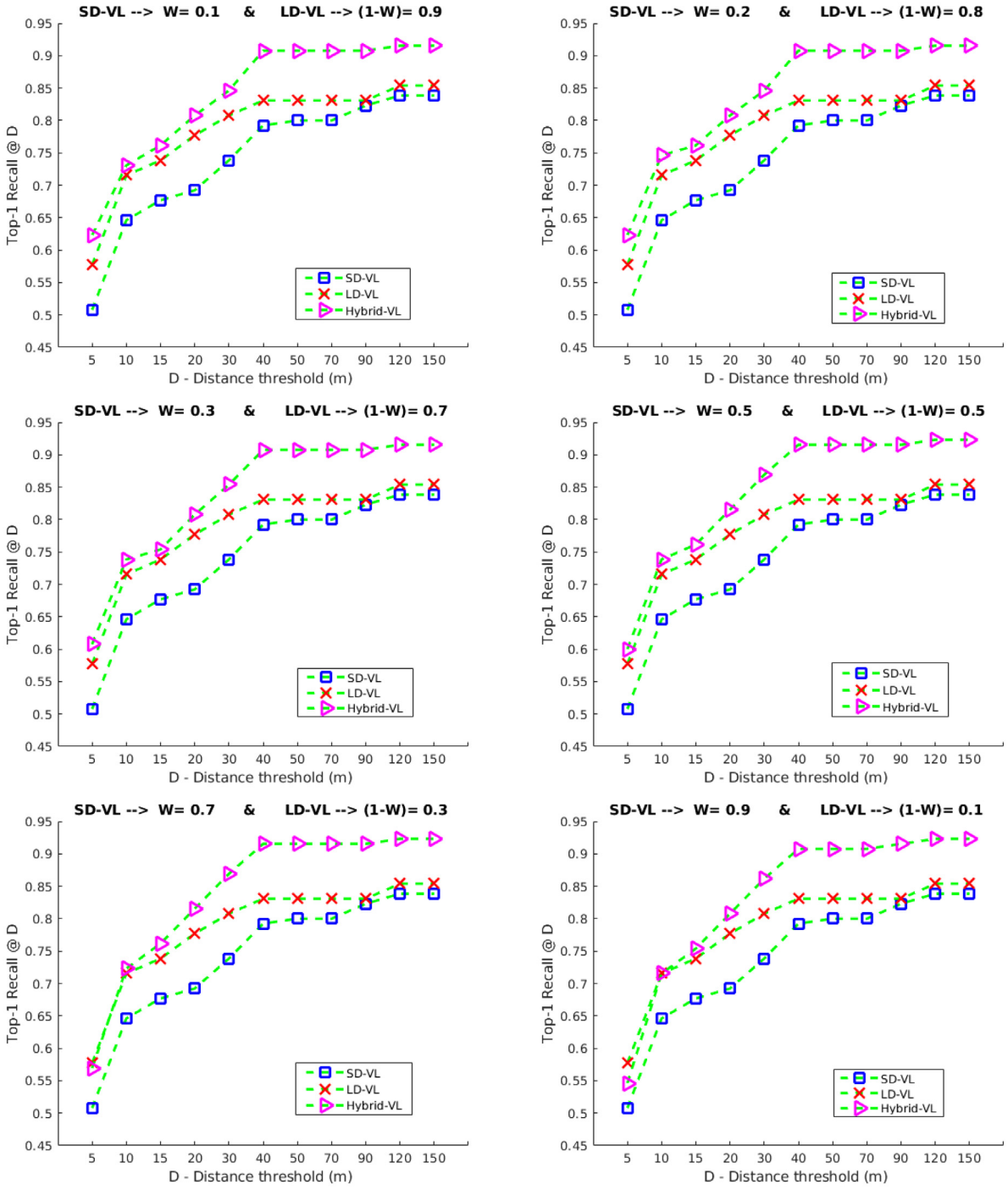

Fig. 12. Effect of altering W parameter method on *RobotCar Seasons* dataset. Best hybridization result (top-right) was obtained with $W = 0.2$, but difference is negligible when compared to $W = 0.1$ and $W = 0.3$.

more weight to SD-VL ($W \geqslant 0.5$) performance of Hybrid-VL decreases.

5. Conclusions and future work

In this study, we propose a Hybrid-VL method that exploits semantic segmentation to improve localization performance. For this purpose, firstly a novel SD is trained with a triplet ranking loss based CNN model using semantically segmented images. Then, this optimized semantic representation is used directly for visual localization named as *learnt* SD-VL method. Lastly, Hybrid-VL method is proposed by combining the newly developed *learnt* SD-VL and the baseline LD-VL methods at decision level.

Improved localization performance is measured with frequently used evaluation metrics on the benchmark *RobotCar Seasons* data set and newly generated *Malaga Streetview Challenge* data set which is shared with the research community. This performance improvement is achieved owing to incorporating the distinguishing power of the relative positions of the objects in a semantically segmented image. We can conclude that the proposed Hybrid-VL method is able to alleviate the shortcomings of the appearance based methods.

As for the future work, employing different kind of descriptors (e.g. using depth maps) would contribute to the success of this work. Furthermore, performing the proposed method on omnidirectional cameras may also increase the localization performance owing to its wide field of viewing angle.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the Scientific and Technological Research Council of Turkey (Grant No.120E500). We also acknowledge the support of NVIDIA Corporation with the donation of Titan Xp GPU used for this research.

References

[1] A. Alzu'bi, A. Abuarqoub, Deep learning model with low-dimensional random projection for large-scale image search, Eng. Sci. Technol., Int. J. 23 (2020) 911–920.
[2] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, L. Van Gool, Night-to-day image translation for retrieval-based localization, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 5958–5964.
[3] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J., 2016. Netvlad: Cnn architecture for weakly supervised place recognition, in: CVPR..
[4] A. Bala, T. Kaur, Local texton xor patterns: A new feature descriptor for content-based image retrieval, Eng. Sci. Technol., Int. J. 19 (2016) 101–112.
[5] J.L. Blanco-Claraco, F.A. Moreno-Duenas, The malaga urban dataset: High-rate stereo and lidars in a realistic urban scenario, Int. J. Robot. Res. 33 (2014).
[6] F. Camposeco, A. Cohen, M. Pollefeys, T. Sattler, Hybrid camera pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 136–144.
[7] Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017a. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587..
[8] Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation, in: ECCV..
[9] Chen, Z., Jacobson, A., Sunderhauf, N., Upcroft, B., Liu, L., Shen, C., Reid, I.D., Milford, M., 2017b. Deep learning features at scale for visual place recognition, in: ICRA..
[10] I. Cinaroglu, Y. Bastanlar, Image based localization using semantic segmentation for autonomous driving, in: 2019 27th Signal Processing and Communications Applications Conference (SIU), IEEE, 2019, pp. 1–4.
[11] I. Cinaroglu, Y. Bastanlar, 23 August, Training semantic descriptors for image-based localization, in: ECCV Workshop on Perception for Autonomous Driving (PAD), 2020.
[12] M. Datar, N. Immorlica, P. Indyk, V.S. Mirrokni, Locality-sensitive hashing scheme based on p-stable distributions, in: Proceedings of the twentieth annual symposium on Computational geometry, 2004, pp. 253–262.
[13] A.D. Doan, Y. Latif, T.J. Chin, Y. Liu, S.F. Ch'ng, T.T. Do, I. Reid, Visual localization under appearance change: filtering approaches, Neural Comput. Appl. (2020) 1–14.
[14] J. Fauqueur, G. Brostow, R. Cipolla, Assisted video object labeling by joint tracking of regions and keypoints, in: 2007 IEEE 11th International Conference on Computer Vision IEEE, 2007, pp. 1–7.
[15] Germain, H., Bourmaud, G., Lepetit, V., 2018. Efficient condition-based representations for long-term visual localization. arXiv preprint arXiv:1812.03707..
[16] L. Huang, M. He, C. Tan, D. Jiang, G. Li, H. Yu, Jointly network image processing: multi-task image semantic segmentation of indoor scene based on cnn, IET Image Proc. 14 (2020) 3689–3697.
[17] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, O. Chum, Panorama to panorama matching for location recognition, in: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, 2017, pp. 392–396.
[18] H. Jegou, M. Douze, C. Schmid, Product quantization for nearest neighbor search, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2010) 117–128.
[19] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, C. Schmid, Aggregating local image descriptors into compact codes, IEEE PAMI 34 (2012) 1704–1716.
[20] D. Jiang, G. Li, C. Tan, L. Huang, Y. Sun, J. Kong, Semantic segmentation for multiscale target based on object recognition using the improved faster-rcnn model, Future Gener. Comput. Syst. 123 (2021) 94–104.
[21] J. Jo, J. Seo, J.D. Fekete, A progressive kd tree for approximate k-nearest neighbors, in: 2017 IEEE Workshop on Data Systems for Interactive Analysis (DSIA), IEEE, 2017, pp. 1–5.
[22] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems (2012) 1097–1105.
[23] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2004) 91–110.
[24] W. Maddern, G. Pascoe, C. Linegar, P. Newman, 1 year, 1000 km: The oxford robotcar dataset, Int. J. Robot. Res. 36 (2017) 3–15.
[25] Mousavian, A., Kosecka, J., Lien, J.M., 2015. Semantically guided location recognition for outdoors scenes, in: ICRA..
[26] M. Muja, D. Lowe, Flann-fast library for approximate nearest neighbors user manual, University of British Columbia, Vancouver, BC, Canada, Computer Science Department, 2009.
[27] M. Muja, D.G. Lowe, Fast approximate nearest neighbors with automatic algorithm configuration, VISAPP 2 (1) (2009) 2.
[28] M. Muja, D.G. Lowe, Scalable nearest neighbor algorithms for high dimensional data, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 2227–2240.
[29] Naseer, T., Oliveira, G.L., Brox, T., Burgard, W., 2017. Semantics-aware visual localization under challenging perceptual conditions, in: ICRA..
[30] S. Orhan, Y. Bastanlar, Efficient search in a panoramic image database for long-term visual localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1727–1734.
[31] Orhan, S., Bastanlar, Y., 2021b. Semantic segmentation of outdoor panoramic images. Signal, Image and Video Processing, 1–8.
[32] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., 2007. Object retrieval with large vocabularies and fast spatial matching, in: CVPR..
[33] N. Piasco, D. Sidibé, C. Demonceaux, V. Gouet-Brunet, A survey on visual-based localization: On the benefit of heterogeneous data, Pattern Recogn. 74 (2018) 90–109.
[34] N. Piasco, D. Sidibé, V. Gouet-Brunet, C. Demonceaux, Learning scene geometry for visual localization in challenging conditions, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 9094–9100.
[35] H. Porav, W. Maddern, P. Newman, Adversarial training for adverse conditions: Robust metric localisation using appearance transfer, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 1011–1018.
[36] F. Radenović, G. Tolias, O. Chum, Fine-tuning cnn image retrieval with no human annotation, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2018) 1655–1668.
[37] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, et al., Benchmarking 6dof outdoor visual localization in changing conditions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8601–8610.
[38] Schönberger, J.L., Pollefeys, M., Geiger, A., Sattler, T., 2018. Semantic visual localization, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition..
[39] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.
[40] Seymour, Z., Sikka, K., Chiu, H.P., Samarasekera, S., Kumar, R., 2018. Semantically-aware attentive neural embeddings for image-based visual localization. arXiv preprint arXiv:1812.03402..
[41] C. Silpa-Anan, R. Hartley, Optimised kd-trees for fast image descriptor matching, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition IEEE, 2008, pp. 1–8.
[42] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556..
[43] Singh, G., Kosecka, J., 2012. Acquiring semantics induced topology in urban environments, in: ICRA..
[44] Sivic, J., Zisserman, A., 2003. Video google: A text retrieval approach to object matching in videos, in: ICCV..
[45] Stenborg, E., Toft, C., Hammarstrand, L., 2018. Long-term visual localization using semantically segmented images, in: (ICRA), pp. 6484–6490..
[46] Sunderhauf, N., Dayoub, F., Shirazi, S., Upcroft, B., Milford, M., 2015. On the performance of convnet features for place recognition, in: IROS..
[47] Toft, C., Stenborg, E., Hammarstrand, L., Brynte, L., Pollefeys, M., Sattler, T., Kahl, F., 2018. Semantic match consistency for long-term visual localization, in: ECCV..
[48] Tolias, G., Sicre, R., Jégou, H., 2015. Particular object retrieval with integral max-pooling of cnn activations. arXiv preprint arXiv:1511.05879..
[49] Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T., 2015a. 24/7 place recognition by view synthesis, in: CVPR..
[50] A. Torii, J. Sivic, M. Okutomi, T. Pajdla, Visual place recognition with repetitive structures, IEEE PAMI 37 (2015).
[51] A. Torii, H. Taira, J. Sivic, M. Pollefeys, M. Okutomi, T. Pajdla, T. Sattler, Are large-scale 3d models really necessary for accurate visual localization?, IEEE Trans Pattern Anal. Mach. (2019). intelligence.
[52] Y. Weiss, A. Torralba, R. Fergus, Spectral hashing, Adv. Neural Inform. Process. Syst. (2009) 1753–1760.
[53] X. Yu, S. Chaturvedi, C. Feng, Y. Taguchi, T.Y. Lee, C. Fernandes, S. Ramalingam, Vlase: Vehicle localization by aggregating semantic edges, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 3196–3203.