2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) | 978-1-6654-8739-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/CVPRW56347.2022.00444

# Semantic Pose Verification for Outdoor Visual Localization with Self-supervised Contrastive Learning

Semih Orhan[1], Jose J. Guerrero[2], Yalin Bastanlar[1]
[1]Department of Computer Engineering, Izmir Institute of Technology
{semihorhan,yalinbastanlar}@iyte.edu.tr
[2]Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza
jguerrer@unizar.es

## Abstract

*Any city-scale visual localization system has to overcome long-term appearance changes, such as varying illumination conditions or seasonal changes between query and database images. Since semantic content is more robust to such changes, we exploit semantic information to improve visual localization. In our scenario, the database consists of gnomonic views generated from panoramic images (e.g. Google Street View) and query images are collected with a standard field-of-view camera at a different time. To improve localization, we check the semantic similarity between query and database images, which is not trivial since the position and viewpoint of the cameras do not exactly match. To learn similarity, we propose training a CNN in a self-supervised fashion with contrastive learning on a dataset of semantically segmented images. With experiments we showed that this semantic similarity estimation approach works better than measuring the similarity at pixel-level. Finally, we used the semantic similarity scores to verify the retrievals obtained by a state-of-the-art visual localization method and observed that contrastive learning-based pose verification increases top-1 recall value to 0.90 which corresponds to a 2% improvement.*

## 1. Introduction

Visual localization can be defined as estimating the position of a visual query material within a known environment. It has received increasing attention [1,7,12,17,28,40] in the last decade especially due to the limitation of GPS-based localization in urban environment (e.g. signal failure in cluttered environment) and motivated by many computer vision application areas such as autonomous vehicle localization [42], unmanned aerial vehicle localization [11], virtual and augmented reality [21].

Visual localization technique that we employ is based on image retrieval, where query images are searched within a geo-tagged database. Location of the retrieved database image serves as the estimated position of the query image. Both query and database images are represented with compact and distinguishable fixed size set of features [1, 3, 12, 29, 30, 44]. In recent years, features extracted with convolutional neural networks (CNNs) [1, 12, 30] outperformed hand-crafted features [2, 17, 27].

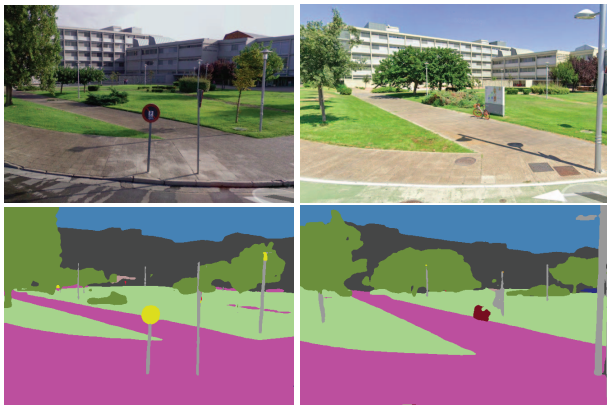

Figure 1. The image on top-left was taken in 2008, the image on top-right was taken in 2019 (source: Google Street View). Observe illumination differences, viewpoint variations and changing objects. Bottom row shows their semantic segmentation results. Semantic similarity can help to verify/deny the localization result.

In our work, query images are collected with a standard field-of-view camera [51], whereas database consists of perspective images (gnomonic projection) generated from a panoramic image dataset (downloaded from Google Street View). The reason for using the panoramic image database is that it presents a wide field-of-view (FOV) which helps to correctly localize the query images where standard FOV cameras fail due to non-overlapping fields of view.

Long-term visual localization remains a challenging research area since the images taken from the environment

978-1-6654-8739-9/22/$31.00 ©2022 IEEE
DOI 10.1109/CVPRW56347.2022.00444

can drastically change over time. Any city-scale visual localization system has to handle appearance changes due to weather conditions, seasonal and illumination variations, as well as structural changes such as building facades. Numerous studies have addressed the aspects of long-term localization in the past [23,25,33,34,36,42,43,45]. Since semantic information is more robust to such changes (Fig. 1), in our study we utilize semantic segmentation as a side modality at the pose verification step. In other words, we check semantic similarity to verify the poses (retrievals) obtained with the approach that use only RGB image features.

To measure semantic similarity, we mainly propose an approach based on self-supervised contrastive learning. We train CNN models [6, 18] using a dataset of unlabeled semantic masks. In our case, unlabeled refers to not knowing if two semantically segmented images belong to the same scene or not. We also have a limited size of labeled dataset (query and database images for the same scene) which is used for fine-tuning and testing purposes. We retrieve top $S$ candidates from the panoramic image database with RGB-only approach and update their similarity scores with semantic features. We showed that, this score update with semantic information improves the performance of a state-of-the-art CNN-based visual localization method (SFRS, [12]).

We also conducted experiments by measuring the semantic similarity at pixel-level, referred to as pixel-wise similarity in the rest of the paper. Since it is the naive approach and easy to implement, we consider it as a baseline. We observed that the pixel-wise similarity can also improve the results of RGB-only approach but when compared to the self-supervised learning with large dataset its gain is marginal.

We summarize our main contributions as follows:

- We adopted self-supervised contrastive learning to represent semantic masks. The trained model is used to estimate similarity scores between the semantic contents of different images.
- Previous visual localization works utilized semantic information for feature point elimination or for performing localization directly by semantic content. In this work, we take a state-of-the-art image-based localization method and improve it using the proposed semantic similarity estimation approach.

The paper is organized as follows. In Section 2, we review the related works. In Section 3, we explain the dataset preparation and demonstrate how we compute and use semantic similarity for pose verification. We present experimental results in Section 4 and conclusions in Section 5.

## 2. Related Work

**Localization with RGB images.** Before the era of CNN, visual localization was mostly performed by representing images with Bag of Visual Words [27], using SIFT-like hand-crafted descriptors [4, 20]. VLAD (Vector of Locally Aggregated Descriptors) [17] does the same task with compact representations that enabled us to use large datasets.

In recent years, CNN-based methods showed great performance on visual localization and image retrieval tasks. One of the first CNN-based approaches was proposed by Razavian *et al.* [31]. They applied the max-pooling function on the last convolution layer of CNN and produced a competitive image representation. Tolias *et al.* [44] improved the previous idea and applied max-pooling to different locations of the convolution layers under different scales. Yi *et al.* [47] proposed the LIFT (Learned Invariant Feature Transform) CNN model that consists of detector, orientation estimator, and descriptor parts. Arandjelovic *et al.* [1] proposed NetVLAD that works on geo-tagged images. The proposed model consists of several convolutions and learnable VLAD layers. SFRS [12], adopted the backbone of NetVLAD and proposed a training regimen to handle the cases with limited overlap. Instead of using database images as a whole during the training, images are divided into parts and similarity scores are calculated on these parts. By doing so, effect of weakness in GPS labels (position errors) was also alleviated and SFRS outperformed previous works on visual localization.

While the majority of localization methods were applied on standard FOV cameras, there are a few previous works on localization with panoramic images [15, 16, 25, 35, 50], but these did not exploit semantic information.

**Semantic-based outdoor localization.** Semantic information is more robust to changes over time and the idea of exploiting semantic content for outdoor visual localization task is not new. We can broadly categorize semantic visual localization methods into 3D structure-based and 2D image retrieval based methods. 3D methods mostly rely on building a 3D model of a scene with structure-from-motion. Stenborg *et al.* [38] performed localization based on the query image's semantic content when the environment is 3D reconstructed and semantically labeled. In another example, 2D-3D point matches are checked if their semantic labels are also matching [43]. In our work, we took the 2D approach which retrieves the most similar image to the query. It arguably performs as well as 3D based methods [34] and less expensive.

Among 2D approaches, Singh and Košecká [37] utilized semantic layout and trained classifier using semantic descriptor to detect intersection points of the streets. Yu *et al.* [48] proposed a method that utilizes semantic edge features (extracted with CASENet [49]). These edges (e.g. sky-building, building-tree) were converted to vector representation and used for localization. Cinaroglu and Bastanlar [8, 9] trained a CNN model with triplet loss on semantic masks and showed that visual localization can solely be done with semantic features. Seymour *et al.* [36] proposed

an attention-based CNN model for 2D visual localization by incorporating appearance and semantic information of the scene. Proposed attention module guide the model to focus on more stable regions. Mousavian *et al.* [22] used semantic information to detect man-made landmark structures (e.g. buildings). Feature points not belonging to man-made objects are considered as unreliable and eliminated. In a similar fashion, Naseer *et al.* [24] applied a weighting scheme on semantic labels (e.g. increasing weights for buildings since they are more stable in long term).

These previous works either performed localization only with semantic labels or they used the semantic features to locate where to focus and eliminate unstable regions. Differently, we check the semantic content in images to validate the retrievals of a state-of-the-art image-based localization method.

**Contrastive learning.** Although its origins date as back as 1990s, contrastive learning has recently gained popularity due to its achievements in self-supervised learning, especially in computer vision [19]. Supervised learning usually requires a decent amount of labeled data, which is not easy to obtain for many applications. With self-supervised learning, we can use inexpensive unlabeled data and achieve a training on a pretext task. Such a training helps us to learn good enough representations. In most cases, a smaller amount of labeled data is used to fine-tune the self-supervised training.

Implemented with Siamese networks, contrastive learning approaches managed to learn powerful representations in a self-supervised fashion. In recently proposed methods, two augmentations of a sample are feed into the networks. The goal of contrastive learning is to learn an embedding space in which similar sample pairs stay close to each other while dissimilar ones are far apart. While MoCo [14] and SimCLR [5] use the negative examples directly along with the positive ones, BYOL [13] and SimSiam [6] achieved similar performance just with the positive examples (different augmentations of the same sample). According to the results, not only image classification, but also object detection and semantic segmentation as downstream tasks benefit from self-supervised contrastive learning.

In our work, we train a CNN model with a contrastive learning approach to learn similarity scores between semantically segmented images. We have a limited size localization dataset (query and database images with known locations), however it is easy to obtain a large dataset of semantic segmentation masks with unknown locations. Thus, we exploited the power of self-supervised learning to learn from a large unlabeled (no location info) dataset. We mainly used SimCLR [5] approach of using both positive and negative samples. To learn similarities in semantic masks, augmented versions of the anchor are taken as positive, samples belong to different scenes are taken as negative (Fig. 2).

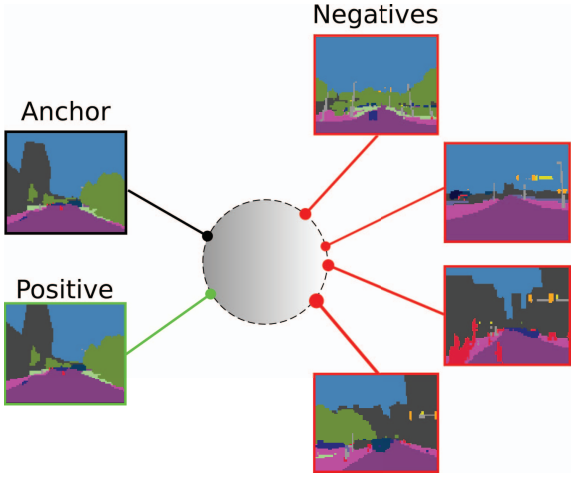


Figure 2. Self-supervised contrastive learning for measuring semantic content similarity. The positive sample is an augmented version of the anchor (we used random crops and small rotations), whereas negative samples belong to different scenes.

## 3. Methodology and Dataset

### 3.1. Dataset

Our dataset consists of images captured in Pittsburgh, PA. Panoramic images in our dataset were obtained from Google Street View (images of 2019) and downloaded with Street View Download 360 application[1]. Query images were taken from UCF dataset [51] at locations corresponding to the panoramic images. These queries were also collected via Google Street View but before 2014. This time gap results in seasonal and structural changes (e.g. change of a facade of a building) in addition to illumination variances. Also, a wide camera baseline between the database and query images conforms better to the long-term localization scenario [23, 33, 43].

To assess the localization performance, we formed a test set consisting of query and database images collected from 123 and 222 different locations respectively. Every query image has a correspondence in the database but not vice versa, i.e. database covers a larger area geographically. Query set consists of 123x4=492 non-overlapping perspective images (90° FOV each). Database consists of 222x12=2674 images (each panoramic image is represented with 12 gnomonic images). Each gnomonic image also has 90° FOV and it overlaps 60° with the next one. Please see Fig.3 for examples.

### 3.2. Searching perspective query images in a panoramic image database

We search query images in the 12-gnomonic image database that is generated from equirectangular panoramic

[1] iStreetView.com

Figure 3. An example set of query and database images taken from the same location. For each location, dataset has one panoramic image (top-left) and four 90° FOV perspective query images (top-right), captured at different times. To localize, we compare each query image with 12 gnomonic images (bottom two rows) generated from the equirectangular panoramic image. Both query and gnomonic database images are 500×400 pixels.

images. Images collected from the same location at different times not only contain appearance changes but also may suffer from limited view overlap due to position and orientation shifts. Even though we use 12 consecutive gnomonic projections on the database side to increase the chance of a good overlap between query and database (there is a 15° viewpoint difference in worst case and 7.5° on average), we can not guarantee a perfect overlap. Several successful methods were proposed to cope with both appearance and viewpoint changes. [1, 12, 44]. Most recently, SFRS [12] outperformed other methods on visual localization benchmark datasets taking advantage of image-to-region similarities. Thus, without loss of generality, we take SFRS [12] as the state-of-the-art image-based method and we verify its retrieval results using semantic similarity.

### 3.3. Computing Semantic Similarity

We first automatically generate a semantic mask for each image in our dataset using a well-performing CNN model [39]. The model we employed was trained on Cityscapes [10], which is an urban scene understanding dataset consists of 30 visual classes, such as building, sky, road, car, etc.

Given a semantic mask, obtaining the most similar result among the alternatives is not a trivial task. SIFT-like features do not exist to match. Moreover, two masks of the same scene is far from being identical not only because of changing content but also due to camera position and viewpoint variations. We have tried geometric methods [32] to fit one image into the other one prior to computing semantic similarity, however they did not succeed. Thus, we propose a trainable semantic feature extractor for pose verification which is trained using correct and incorrect pose matchings.

Before presenting the trainable approach, we explain pixel-wise similarity approach which measures the semantic similarity at pixel-level. Since it is the naive approach and easy to implement, we see this as a baseline method of semantic pose verification.

**Pixel-wise Similarity.** In this first approach, we calculate pixel-by-pixel similarity between query and database semantic masks:

$$\textit{pixel-wise similarity} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} sim(Q_{(i,j)}, D_{(i,j)})}{m \cdot n} \quad (1)$$

where $sim(a, b)$ is equal to 1 if $a = b$, 0 otherwise. $Q$ represents the query image's mask and $D$ represents the database image's mask, both having size $m \times n$, $(i, j) \in \{1, ..., m\} \times \{1, ..., n\}$. A pixel is considered as a matching pixel if $Q_{(i,j)} = D_{(i,j)}$ and it increases similarity.

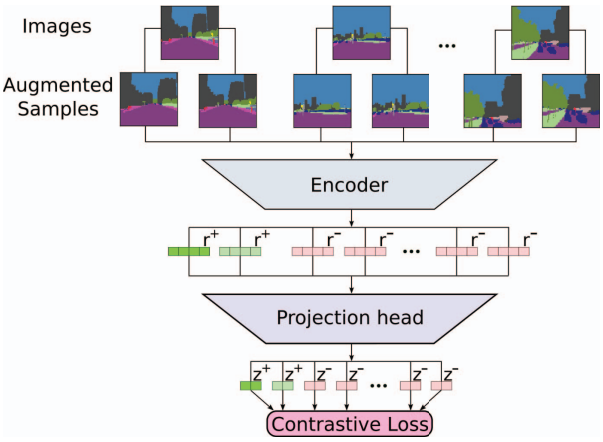


Figure 4. Illustration of training a CNN model with self-supervised contrastive loss on a dataset that consists of semantically segmented masks.

**Trainable Feature Extractor.** We use self-supervised contrastive learning approach in our work since large amount of semantic masks can easily be obtained for a self-supervised training. In our setting, semantic masks are obtained with a well-performing segmentation model [39] from 3484 images randomly taken from UCF dataset [51]. We do not need groundtruth masks, since a successful estimation is enough to compute semantic similarity.

We used SimCLR [5] as our contrastive learning model and trained a ResNet-18 as the encoder. In our setting, encoder network (Fig. 4) produces $r = Enc(x) \in R^{512}$ dimensional features, projection network produces $z = Proj(r) \in R^{2048}$ dimensional features. We resized semantic mask to $64 \times 80$ resolution (due to GPU memory limitation) and used two different data augmentation methods during the training: random resized crop and random rotation. We set lower bound of random crop parameter as 0.6, which means that cropped mask covers at least 60% area of the original mask. We set maximum rotation parameter as 3°, since severe rotations are not expected between query and database images. Augmentation of semantic masks is visualized in Fig. 4. Following [5,18] we use the contrastive loss given in Eq. 2. This is a categorical cross-entropy loss to identify the positive sample amongst a set of negative samples (inspired from InfoNCE [46]).

$$L^{self} = \sum_{i \in I} L_i^{self} = -\sum_{i \in I} \log \frac{\exp(z_i \cdot z_{j(i)}/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_{a(i)}/\tau)} \tag{2}$$

$N$ images are randomly taken from the dataset. Thus, the training batch consists of $2N$ images to which data augmentations are randomly applied. Let $i \in I \equiv \{1...2N\}$ be the index of an arbitrary augmented sample, then $j(i)$ is the index of the other augmentation of the same original image. $\tau \in R^+$ is a scalar temperature parameter, $\cdot$ represents the dot product, and $A(i) \equiv I - \{i\}$. We call index $i$ the anchor, index $j(i)$ is the positive, and the other $2(N-1)$ indices as negatives. The denominator has a total of $2N-1$ terms (one positive and $2N-2$ negatives).

CNN model, trained as explained above, is now ready to produce a similarity score when two semantic masks (one query and one database) are given. Similarity score is used to update the scores of RGB-only method (Section 3.4).

After self-supervised training, same network can be fine-tuned with a labeled dataset (query and database segmentation masks for the same scene). For this purpose, we prepared a dataset of 227 query images with their corresponding database panoramic images. Not surprisingly, it is much smaller than the self-supervised training dataset. Here, common practice in literature is that the projection head (Fig.4) is removed after pretraining and a classifier head is added and trained with the labeled data for the downstream task. However, since our pretraining and downstream tasks are the same (estimating similarity of two input semantic masks), we do not place a classifier head, but we retrain the network (partially or full).

### 3.4. Updating Retrieval Results with Semantic Similarity

We first normalize RGB-only [12] and semantic similarity (pixel-wise similarity or trainable feature extractor) scores between $[-1,+1]$ and then merge them with a weight coefficient ($W$) to obtain the updated similarity score:

$$\textit{updated-score}_i = \textit{rgb-score}_i + W \cdot \textit{semantic-score}_i \tag{3}$$

where $i$ is the index within the top $S$ candidates for each query. We do not update the scores of every database image, but only interested in the top $S$ database candidates for each query, since these are already obtained by a state-of-the-art image-based localization method. We set $S$=10 in our experiments. A real (successful) example of similarity score update is shown in Fig.5.

While updating the similarity scores of $S$ candidates, we also update the similarity scores of gnomonic images coming from the same panorama (neighbours of the retrieved gnomonic views also have potential to benefit from semantic similarity).

To decide on $W$ value, we employed a validation set of query-database image pairs. We selected $W$ values with the highest localization performance, separately for pixel-wise similarity and trainable semantic feature extractor approaches. Effect of altering $W$ was examined with experiments in Section 4.1.

## 4. Experimental Results

Our self-supervised contrastive learning uses 3484 images randomly taken from UCF dataset [51] and not co-

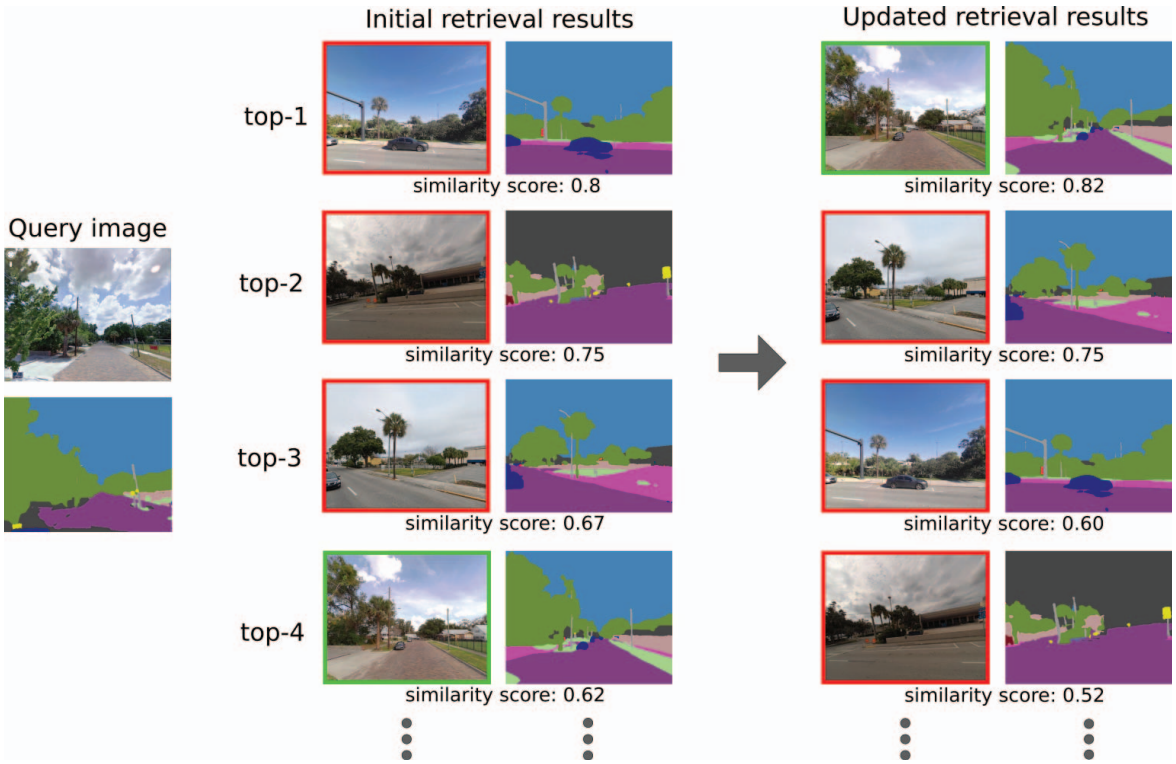


Figure 5. An example set of retrieval results from the database where RGB-only method fails to correctly localize but the method merging RGB and semantic scores correctly localizes. Green rectangle indicates correct match, which increased to top rank after score update.

incide with the localization test set. We used stochastic gradient descent optimizer with initial learning rate = 0.05. Temperature parameter ($\tau$) was taken as 0.07 and batch size ($2N$) as 174.

With experiments, we compare the localization performances of three approaches on the test set explained in Section 3.1 (492 query and 2674 database images). First approach is the state-of-the-art visual localization with RGB image features [12] without additional training, second approach is updating RGB-only method's scores with pixel-wise semantic similarity, and third is updating RGB-only method's scores with the similarity given by trainable semantic feature extractor (with SimCLR [5] self-supervised training scheme). We also conducted experiments with models fine-tuned with labeled dataset after self-supervised training. Those results will be presented in Section 4.1.

Performances of different approaches are compared with Recall@N metric. According to this metric, a query image is considered as correctly localized if the distance between the query and any retrieved database images in top-N is smaller than the metric distance threshold. The threshold was set as 5 meters in our experiments. Since the test dataset is prepared so that a database image is taken at the location of each query image, ideally all the queries can be localized with 5-meter threshold. Results in Fig.7 show that the semantic pose verification is useful for all cases and it improves Recall@1 of RGB-only model by %2 when the proposed trainable semantic feature extractor is trained in a self-supervised fashion with SimCLR. In addition to Recall@N plots in Fig.7, we provide Recall@1 performances with varying distance thresholds in Fig.8. We observe that pose verification with trainable feature extractor continues to outperform RGB-only approach and pose verification with pixel-wise similarity approach for increasing distance thresholds. Lastly, Fig.6 shows several examples where the proposed semantic pose verification improved the results.

### 4.1. Ablation Study

Results given so far were obtained with self-supervised contrastive learning without any fine-tuning, minimum crop ratio was used as 0.6 and $W$ was set to 0.25 as suggested values. Now, we present our fine-tuning results and additional experiments to investigate how much our approach is sensitive to crop ratio and $W$ parameters.

Table 1 compares self-supervised learning results with alternatives where the model is fine-tuned with a labeled dataset which corresponds to 227 query semantic masks and their actual corresponding semantic masks in the database. Three different fine-tuning schemes were tested: last two dense layers were retrained, two new dense layers were

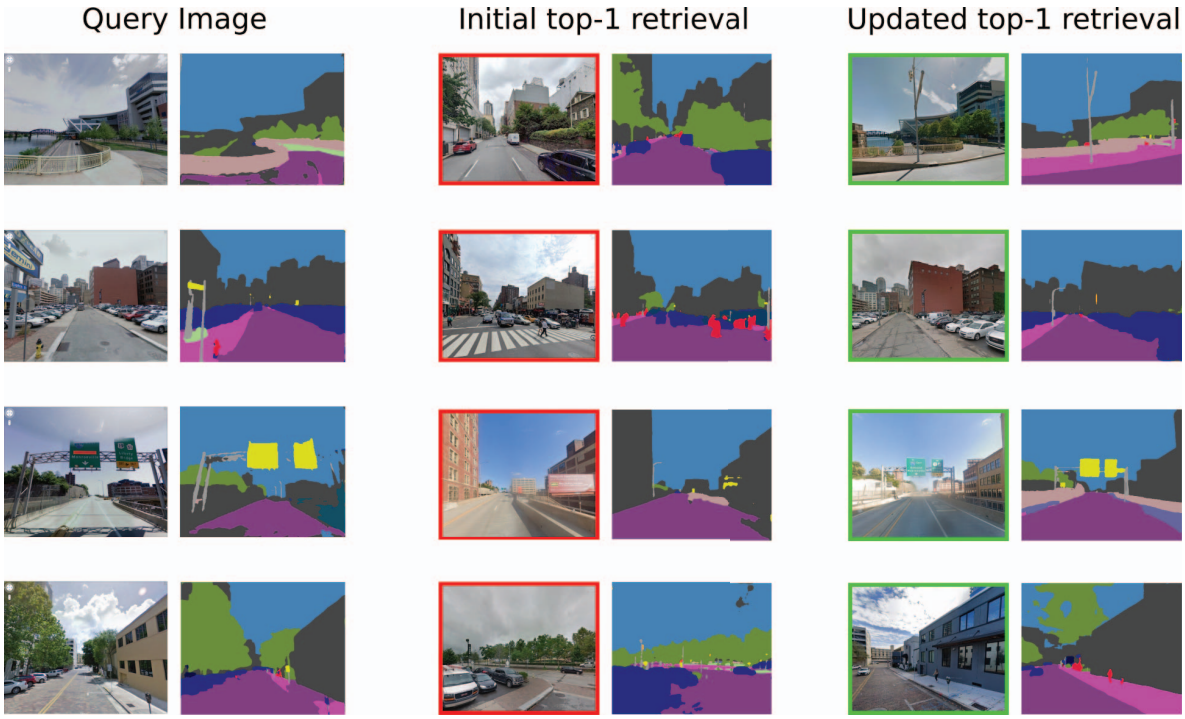


Figure 6. Example retrieval results in which utilizing semantic similarity scores at pose verification improved the localization performance of the RGB-only model. Query images are in the first column, top-1 retrieval results are in the middle column, and updated top-1 retrieval results with trainable semantic feature extractor are presented in the last column. Utilizing semantic similarity moved up the correct candidates in ranking when semantic contents of query and database images are similar. Distinctive objects (e.g. traffic signs) help to correctly localize query images with semantic information (third row of the figure). In some cases, localization was improved even the semantic masks of the images contain labeling errors (last row of the figure).

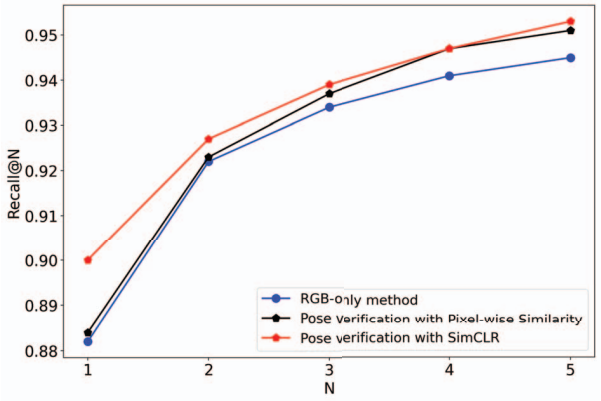


Figure 7. Visual-based localization results of RGB-only method, pose verification with pixel-wise similarity, and pose verification with self-supervised learning (SimCLR).

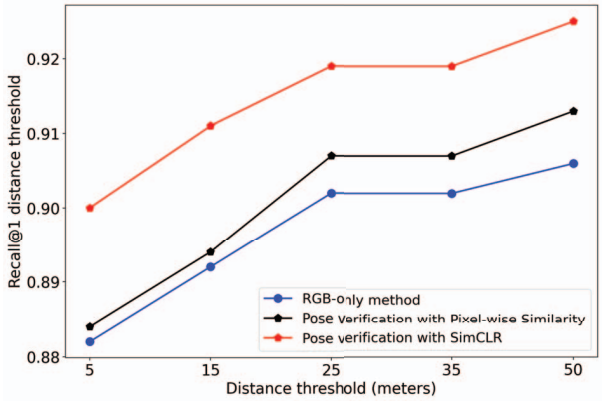


Figure 8. Localization results of three approaches where N=1 with different distance thresholds.

added and trained, all layers of network were trained. None of them improved self-supervised training and only fine-tuning all layers kept a similar performance. It should be noted that our labeled dataset is much smaller than the unlabeled dataset (227 $\ll$ 3484). Another reason for not improving with fine-tuning could be the fact that our main and downstream tasks are the same, i.e. scoring similarity between two semantic masks. Whereas, successful examples of fine-tuning in literature contains placing a classifier head and training for a different downstream task like image classification or object detection.

Table 1. Self-supervised model is compared with fine-tuned models. Results were obtained with semantic score weight $W = 0.25$.

| Training Methods | Recall@N | | | | |
|---|---|---|---|---|---|
| | N=1 | N=2 | N=3 | N=4 | N=5 |
| Only self-supervised training | **0.900** | 0.927 | 0.939 | 0.947 | 0.953 |
| Fine-tuning last two dense layers | 0.890 | 0.927 | 0.937 | 0.943 | 0.945 |
| Adding two new dense layers | 0.892 | 0.927 | 0.935 | 0.947 | 0.947 |
| Fine-tuning all layers | **0.900** | 0.925 | 0.933 | 0.943 | 0.945 |

Table 2. Effect of the minimum crop ratio parameter in data augmentation on localization performance.

| Crop Ratio | Recall@N | | | | |
|---|---|---|---|---|---|
| | N=1 | N=2 | N=3 | N=4 | N=5 |
| 0.90 | 0.888 | 0.921 | 0.931 | 0.939 | 0.939 |
| 0.80 | 0.892 | 0.927 | 0.937 | 0.943 | 0.947 |
| 0.70 | 0.896 | 0.931 | 0.941 | 0.943 | 0.947 |
| 0.60 | **0.900** | 0.927 | 0.939 | 0.947 | 0.953 |
| 0.50 | 0.898 | 0.929 | 0.937 | 0.941 | 0.943 |
| 0.40 | 0.894 | 0.931 | 0.937 | 0.943 | 0.947 |
| 0.30 | 0.894 | 0.927 | 0.933 | 0.943 | 0.947 |

Table 3. Effect of the weight of semantic similarity score ($W$) on localization performance.

| Semantic Weight | Recall@N | | | | |
|---|---|---|---|---|---|
| | N=1 | N=2 | N=3 | N=4 | N=5 |
| 0.10 | 0.884 | 0.925 | 0.935 | 0.945 | 0.951 |
| 0.15 | 0.884 | 0.929 | 0.937 | 0.943 | 0.951 |
| 0.20 | 0.892 | 0.929 | 0.937 | 0.949 | 0.953 |
| 0.25 | **0.900** | 0.927 | 0.939 | 0.947 | 0.953 |
| 0.30 | **0.902** | 0.929 | 0.937 | 0.945 | 0.953 |
| 0.35 | **0.902** | 0.931 | 0.937 | 0.943 | 0.949 |
| 0.40 | **0.900** | 0.929 | 0.937 | 0.945 | 0.949 |
| 0.45 | 0.888 | 0.925 | 0.939 | 0.945 | 0.947 |
| 0.50 | 0.882 | 0.923 | 0.939 | 0.945 | 0.947 |

We have also evaluated another self-supervised training approach, SimSiam [6], however its performance was worse than SimCLR. Thus, we excluded it from our ablation study.

Table 2 presents the effect of minimum crop ratio parameter used in data augmentation module. Values fluctuate in a close range for N=$\{2,..,5\}$, however, Recall@1 is highest for $0.6$ and performance gradually drops as we increase or decrease the minimum crop ratio. This is in accordance with the finding in [41] that there is a reverse-U shaped relationship between the performance and the mutual information within augmented views. When crops are close to each other (high mutual information, e.g. crop ratio=$0.9$) the model does not benefit from them much. On the other hand, for low crop ratios (low mutual information) model can not learn well since views look quite different from each other. Peak performance stays somewhere in between.

Lastly, Table 3 shows the results of the experiments with different semantic weight coefficient ($W$). We understand that success is not specific to $W = 0.25$ and pose verification works equally well for values between $0.25$ and $0.40$.

## 5. Conclusion

In this work, we localize perspective query images in a geo-tagged database of panoramic images. We take advantage of semantic segmentation masks due to their robustness to long-term changes. Semantic similarity is measured via pixel-wise similarity and trainable feature extractors. Experimental results showed that utilizing semantic similarity at pose verification step contributed to visual localization performance of a state-of-the-art method [12]. Gained improvement is due to the more stable semantic content and does not depend on which localization method used to obtain initial retrieval results. Thus, other RGB image based visual localization methods can be improved in the same manner.

We also conclude that pose verification with a CNN model, which exploits self-supervised contrastive learning, performs better than using pixel-wise similarity between masks. This confirms the potential of self-supervised models for representation learning when there is a limited amount of labeled data.

There are works that search the query image within the panoramic image instead of using gnomonic views (e.g. [16, 25]). Also, semantic segmentation masks can be obtained for panoramic images of street views [26]. A future work may be extending our effort to compute semantic similarity directly from panoramic semantic masks.

## Acknowledgments

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under Grant No. 120E500 and also under 2214-A International Researcher Fellowship Programme.

## References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Net VLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016. 1, 2, 4

[2] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2911–2918. IEEE, 2012. 1

[3] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. From generic to specific deep representations for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 36–45, 2015. 1

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision*, pages 404–417. Springer, 2006. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607, 2020. 3, 5, 6

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2, 3, 8

[7] Zetao Chen, Adam Jacobson, Niko Sünderhauf, Ben Upcroft, Lingqiao Liu, Chunhua Shen, Ian Reid, and Michael Milford. Deep learning features at scale for visual place recognition. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3223–3230, 2017. 1

[8] I. Cinaroglu and Y. Bastanlar. Training semantic descriptors for image-based localization. In *ECCV Workshop on Perception for Autonomous Driving*, 2020. 2

[9] I. Cinaroglu and Y. Bastanlar. Long-term image-based vehicle localization improved with learnt semantic descriptors. *Engineering Science and Technology, an International Journal (JESTECH)*, 35, 2022. 2

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 4

[11] Andy Couturier and Moulay A. Akhloufi. A review on absolute visual localization for UAV. *Robotics and Autonomous Systems*, 135:103666, 2021. 1

[12] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In *European Conference on Computer Vision*, pages 369–386, 2020. 1, 2, 4, 5, 6, 8

[13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. 3

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[15] Jiung-Yao Huang, Su-Hui Lee, and Chung-Hsien Tsai. A fast image matching technique for the panoramic-based localization. In *IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, 2016. 2

[16] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, and O. Chum. Panorama to panorama matching for location recognition. In *ACM International Conference on Multimedia Retrieval (ICMR)*, 2017. 2, 8

[17] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, 2011. 1, 2

[18] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 2, 5

[19] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 2020. 3

[20] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 2

[21] Simon Lynen, Bernhard Zeisl, Dror Aiger, Michael Bosse, Joel Hesch, Marc Pollefeys, Roland Siegwart, and Torsten Sattler. Large-scale, real-time visual–inertial localization revisited. *The International Journal of Robotics Research*, 39(9):1061–1084, 2020. 1

[22] Arsalan Mousavian, Jana Košecká, and Jyh-Ming Lien. Semantically guided location recognition for outdoors scenes. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4882–4889. IEEE, 2015. 3

[23] Y. Naiming, T. Kanji, F. Yichu, F. Xiaoxiao, I. Kazunori, and I. Yuuki. Long-term vehicle localization using compressed visual experiences. In *21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018. 2, 3

[24] Tayyab Naseer, Gabriel L Oliveira, Thomas Brox, and Wolfram Burgard. Semantics-aware visual localization under challenging perceptual conditions. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2614–2620. IEEE, 2017. 3

[25] Semih Orhan and Yalin Bastanlar. Efficient search in a panoramic image database for long-term visual localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 2, 8

[26] Semih Orhan and Yalin Bastanlar. Semantic segmentation of outdoor panoramic images. *Signal, Image and Video Processing*, 2021. 8

[27] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 1, 2

[28] Nathan Piasco, Desire Sidibe, Cedric Demonceaux, and Gouet-Brunet Valerie. A survey on visual-based localization: On the benefit of heterogeneous data. *Pattern Recognition*, 74:90–109, 2018. 1

[29] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings*

*of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5706–5715, 2018. 1

[30] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2018. 1

[31] Ali S Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications*, 4(3):251–258, 2016. 2

[32] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 4

[33] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DOF outdoor visual localization in changing conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8601–8610, 2018. 2, 3

[34] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. Are large-scale 3d models really necessary for accurate visual localization? In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6175–6184, 2017. 2

[35] Georg Schroth, Robert Huitl, David Chen, Mohammad Abu-Alqumsan, Anas Al-Nuaimi, and Eckehard Steinbach. Mobile visual location recognition. *IEEE Signal Processing Magazine*, 28(4):77–89, 2011. 2

[36] Zachary Seymour, Karan Sikka, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. Semantically-aware attentive neural embeddings for long-term 2d visual localization. In *British Machine Vision Conference*, 2019. 2

[37] Gautam Singh and Jana Košecká. Acquiring semantics induced topology in urban environments. In *2012 IEEE International Conference on Robotics and Automation*, pages 3509–3514. IEEE, 2012. 2

[38] Erik Stenborg, Carl Toft, and Lars Hammarstrand. Long-term visual localization using semantically segmented images. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6484–6490. IEEE, 2018. 2

[39] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019. 4, 5

[40] Hajime Taira, Ignacio Rocco, Jiri Sedlar, Masatoshi Okutomi, Josef Sivic, Tomas Pajdla, Torsten Sattler, and Akihiko Torii. Is this the right place? geometric-semantic pose verification for indoor visual localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4373–4383, 2019. 1

[41] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 8

[42] Carl Toft, Will Maddern, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, et al. Long-term visual localization revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2

[43] Carl Toft, Erik Stenborg, Lars Hammarstrand, Lucas Brynte, Marc Pollefeys, Torsten Sattler, and Fredrik Kahl. Semantic match consistency for long-term visual localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 383–399, 2018. 2, 3

[44] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of CNN activations. *International Conference on Learning Representations (ICLR)*, 2016. 1, 2, 4

[45] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2015. 2

[46] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[47] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European Conference on Computer Vision*, pages 467–483, 2016. 2

[48] Xin Yu, Sagar Chaturvedi, Chen Feng, Yuichi Taguchi, Teng-Yok Lee, Clinton Fernandes, and Srikumar Ramalingam. VLASE: Vehicle localization by aggregating semantic edges. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3196–3203. IEEE, 2018. 2

[49] Zhiding Yu, Chen Feng, Ming-Yu Liu, and Srikumar Ramalingam. Casenet: Deep category-aware semantic edge detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5964–5973, 2017. 2

[50] Amir Roshan Zamir and Mubarak Shah. Accurate image localization based on google maps street view. In *European Conference on Computer Vision*, pages 255–268. Springer, 2010. 2

[51] Amir Roshan Zamir and Mubarak Shah. Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1546–1558, 2014. 1, 3, 5