

Hierarchical localization with panoramic views and triplet loss functions

Marcos Alfaro^{a,*}, Juan José Cabrera^a, Luis Miguel Jiménez^a, Óscar Reinoso^a, Luis Payá^a

^a*Miguel Hernández University, Avenida de la Universidad
s/n, Elche, 03202, Comunidad Valenciana, Spain*

Abstract

The main objective of this paper is to address the mobile robot localization problem with Triplet Convolutional Neural Networks and test their robustness against changes of the lighting conditions. We have used omnidirectional images from real indoor environments captured in dynamic conditions that have been converted to panoramic format. Two approaches are proposed to address localization by means of triplet neural networks. First, hierarchical localization, which consists in estimating the robot position in two stages: a coarse localization, which involves a room retrieval task, and a fine localization is addressed by means of image retrieval in the previously selected room. Second, global localization, which consists in estimating the position of the robot inside the entire map in a unique step. Besides, an exhaustive study of the loss function influence on the network learning process has been made. The experimental section proves that triplet neural networks are an efficient and robust tool to address the localization of mobile robots in indoor environments, considering real operation conditions.

Keywords: robot localization, convolutional neural network, omnidirectional images, triplet loss

*Corresponding author: Marcos Alfaro

Email addresses: marcos.alfaro@goumh.umh.es (Marcos Alfaro), juan.cabreram@umh.es (Juan José Cabrera), luis.jimenez@umh.es (Luis Miguel Jiménez), o.reinoso@umh.es (Óscar Reinoso), lpaya@umh.es (Luis Payá)

1. Introduction

Nowadays, vision sensors are frequently used to address the localization problem in mobile robotics, since they can capture a large amount of information from the environment at a low cost. Among these sensors, omnidirectional cameras stand out (Amorós *et al.* [1]). This type of cameras have a field of view up to 360° , so they capture complete information from the environment regardless of the robot orientation. Omnidirectional views can be obtained with different alternatives, such as multicamera systems (Kneip *et al.* [17]), catadioptric systems (Lin *et al.* [23]) or the combination of a pair of fisheye cameras (Flores *et al.* [12]).

In order to describe the visual information from the scene, two main approaches have been proposed in the related work. First, holistic or global description consists in working with the image information as a whole (Payá *et al.* [31]), whereas the description based on local features only focuses on those points or areas easily identifiable in an image, such as borders or corners (Murillo *et al.* [27]). In this work, global description is used.

Traditionally, analytical techniques have been used to create visual descriptors (Se *et al.* [36]). However, with the huge increase of computing power, the use of deep learning tools has increased during the past few years. Concerning to image processing, Convolutional Neural Networks (CNNs) are the most extended approach (Nilwong *et al.* [28], Cebollada *et al.* [5]). This type of neural networks apply filters to the image based on the convolution operation, and are able to extract features from the image with a high level of abstraction.

In recent years, other works have explored the use of more complex architectures, composed of several neural networks, giving place to Siamese Networks (Yin *et al.* [46]) and Triplet Networks (Liu and Huang [25]), among others. Siamese networks contain two identical neural networks, that is, they have the same architecture and share their weights, and work in parallel, in such a way that each of them receives a different input and provides a different output. Meanwhile, Triplet Networks receive three inputs, commonly called anchor, positive and negative, and provide three outputs. While Siamese Networks are typically used to learn if two inputs are similar or different, Triplet Networks are able to simultaneously learn similarities between the

anchor and positive inputs and differences between the anchor and negative data.

During the training process, the loss function compares the output provided by the network with the required output, and the optimization of this function leads to more accurate predictions. As a function of the loss value, the optimizer algorithm modifies the network weights to a greater or smaller extent. Triplet loss functions (Hermans *et al.* [14]) seek to minimize the difference between the anchor and positive inputs and also seek to maximize the difference between the anchor and negative inputs. This type of loss functions have some parameters that must be set before the training. The most relevant is the margin, which permits adjusting the required similarity and difference relationships between the data.

In this paper, a convolutional neural network model is used, which is adapted and retrained to tackle the localization of a mobile robot in indoor environments with panoramic images, employing a triplet network architecture. The experimental section shows the robustness of such architecture to address localization. Thanks to it, a scarce training with a limited set of images captured under a specific lighting condition is enough to obtain a tool which is robust against changes in the lighting conditions and capable of adapting to different environments without the need of a data augmentation process. In addition, an exhaustive comparative evaluation between several triplet losses has been performed at every localization stage.

Therefore, the main contributions of the present work are:

- A hierarchical localization approach which exploits the advantages of triplet network architectures in indoor environments is proposed.
- Triplet networks are trained and evaluated with panoramic images, obtained from a catadioptric system mounted on a mobile robot. Besides, their robustness is analyzed against defying visual phenomena such as lighting changes or visual aliasing.
- We conduct a complete comparative evaluation of the performance of different triplet loss functions in the global and the hierarchical localization.

The manuscript is structured as follows. Section 2 reviews the state of the art on robot localization, holistic visual description and the use of deep learning to perform these tasks. Section 3 presents the network architecture and the loss functions used in this work. In section 4, the two localization methods employed in this paper are detailed. Section 5 describes the experiments conducted. Finally, in section 6 the conclusions and future works are outlined.

2. State of the art

Nowadays, the use of vision systems in mobile robotics is very common. Many research works make use of cameras to solve the localization and mapping problems. Among this type of sensors, monocular cameras are the most extended option. For example, Xiao *et al.* [44] addressed the SLAM problem in dynamic environments with a monocular vision system. Other works make use of omnidirectional vision systems as they can capture complete information from the scenario regardless of the robot orientation. Flores *et al.* [12] perform localization with omnidirectional and fisheye cameras.

With respect to visual description, there are some authors, such as Payá *et al.* [31] or Cebollada *et al.* [6], that propose environment modeling techniques with global-appearance descriptors. Moreover, some works make use of these descriptors to tackle the loop closure problem, one of the most critical parts of SLAM algorithms (Zhang *et al.* [48]). Also, local descriptors are commonly used as well to perform localization (Kallasi *et al.* [15]). Furthermore, other works combine the two types of descriptors to address mapping and/or localization (Li *et al.* [22], Su *et al.* [38]).

The increase of computing power has led to the rise of Convolutional Neural Networks in the past decade. When it comes to process visual information captured by a robot, this type of networks proved to be able to extract features from the image and therefore solve mobile robotics problems like visual localization. CNNs were first proposed in [20], and further developed in subsequent works, which propose more complex architectures, such as VGG (Simonyan and Zisserman [37]), GoogLeNet (Szegedy *et al.* [40]) or AlexNet (Krizhevsky *et al.* [19]), all of them trained to classify a thousand different objects with the ImageNet database (Deng *et al.* [10]). Although CNNs are the most extended choice, lately other architectures have been pro-

posed to process visual information. This is the case of Visual Transformers (Dosovitskiy *et al.* [11]), which are based on Transformers, commonly used in Natural Language Processing. Besides, other works propose different networks that are able to process 3D point clouds (Qi *et al.* [33], Komorowski [18]).

Focusing on CNNs, many recent works use them to address visual localization. For instance, Nilwong *et al.* [28] make use of local features obtained with a CNN from RGB images captured in outdoor environments, and Foroughi *et al.* [13] did the same indoors. Others, such as Xu *et al.* [45], make use of feature descriptors extracted from different convolutional layers of the network. CNNs can also be trained to obtain global-appearance descriptors from the image (Cabrera *et al.* [4]). Moreover, Chen *et al.* [8] propose a two-step method by combining global and local features. First, an image retrieval phase takes place by comparing global image descriptors. Second, the robot pose is estimated by comparing the ORB keypoints of the captured image with the keypoints in the two most similar images. Rostkowska and Skrzypezynski [35], Ballesta *et al.* [3] and Cebollada *et al.* [5] also perform a hierarchical localization by identifying in first place the room where the robot has captured the image and later estimate the robot coordinates inside the room predicted in the first step. Besides, Wozniak *et al.* [43] train a CNN to classify images among 16 rooms.

Due to the success of CNNs, other works have implemented advanced architectures composed of several CNNs. Siamese Networks are composed of two identical neural networks that work in parallel and share their weights. Apart from being able to extract global features from the image, siamese networks can include some additional layers to evaluate the similarity between the two inputs. This ability can be used in mobile robotics tasks such as place recognition (Leyva-Vallina *et al.* [21]), loop closure (Qiu *et al.* [34]) or visual localization (Oliveira *et al.* [30]). Other researchers have designed a siamese architecture that is not composed of CNNs. For example, Chen *et al.* [7] make use of a siamese network to evaluate LiDAR scan similarity. Each network receives a LiDAR 3D point cloud and embeds the representation into the euclidean space to estimate their similarity.

Likewise, Triplet Networks contain three identical networks. These architectures receive three inputs, called anchor, positive and negative, and

provide three different outputs. Triplet networks are trained with combinations of three images, and in the case of robot localization, they can be chosen in such a way that two of them are captured from similar positions and the other is captured from a different position. The fact of receiving three inputs permits the network to adjust both to positive and negative examples during the training process. Besides, since the number of possible combinations of three images is very large, a fairly small number of images captured by the robot can be enough to create a complete training set. Even though, triplet networks have barely been used in visual localization tasks, and only few approaches can be found in recent years. Also, all of them used standard cameras or RGB-d cameras. Arandjelovic *et al.* [2] designed a triplet network that aggregates the extracted local features into a single descriptor using a VLAD layer. Yu *et al.* [47] also make use of a VLAD layer to address the same problem. López-Antequera *et al.* [26] proposed a triplet network architecture to carry out a visual localization under seasonal changes. Likewise, Olid *et al.* [29] make a comparative evaluation of several CNN, siamese and triplet networks, obtaining the highest recall with triplet architectures. Comparing to these works, in the present work we propose a hierarchical localization approach, which exploits the advantages of the triplet networks in challenging indoor environments. Also, we explore the use of triplet networks along with panoramic images, obtained from a catadioptric system mounted on the robot both to train and test the architectures.

The development of triplet networks goes hand in hand with the design of triplet loss functions. Some works have focused on creating a loss function that optimizes the training of their triplet architecture. Hermans *et al.* [14] compare different triplet loss functions used to train a network for people recognition. Cheng *et al.* [9] use a variant of the Triplet Margin Loss, proposed in [14], to solve the same problem. Nevertheless, there have been only few works that designed a triplet loss function to tackle visual localization. For example, Liu *et al.* [24] created a triplet loss function and compared it with other loss functions to solve a place recognition problem. Also, Kim *et al.* [16] developed a triplet loss function to undertake a room retrieval task. Even though, triplet loss functions have not been thoroughly tested in visual localization tasks and in the present work we perform a complete comparative evaluation of the performance of such loss functions and the influence of their parameters in the global and hierarchical localization with panoramic images.

3. Architecture of the neural network and triplet losses

Triplet Neural Networks consist of three identical neural networks that work in parallel and share their weights, but each of them can receive a different input and therefore will provide a different output. These structures are trained with combinations of three input data vectors, commonly called anchor, positive and negative. The network is trained to learn similarities between the anchor and positive vectors and differences between the negative vector and the other two inputs. In some applications, Triplet Networks present some advantages over Siamese Networks, which are composed of a pair of Neural Networks. First, Triplet Networks receive the same number of positive and negative inputs, which allows the network to adjust equally to similar and different data during the training process. This property can be especially useful in localization tasks, especially in those indoor environments which are prone to visual aliasing. Second, the number of possible input combinations in the training process increases substantially compared with Siamese Networks. This can be especially useful when just a scarce dataset is initially available, because a reasonably high number of triplet samples can be obtained to train the network even if no data augmentation is performed. For these reasons, triplet neural networks can play a remarkable role to solve the visual localization of a mobile robot, and we address this problem in the present work.

In order to carry out localization by using a triplet convolutional network architecture, we make use of the VGG16 network model [37], which is adapted as shown in the Figure 1. In first place, given that the size of the panoramic images in this work is 128x512x3 pixels, the first fully connected layer of the feature aggregation stage must be adapted to this size (its original size was 224x224x3 pixels). Additionally, we leave the convolutional layers intact, which correspond to the feature extraction phase, and modify the remaining fully connected layers so as to obtain a five-element global-appearance descriptor, as shown in the Figure 1. With the aim of taking advantage of the knowledge already acquired by the VGG16 model, Transfer Learning technique is employed on the convolutional layers.

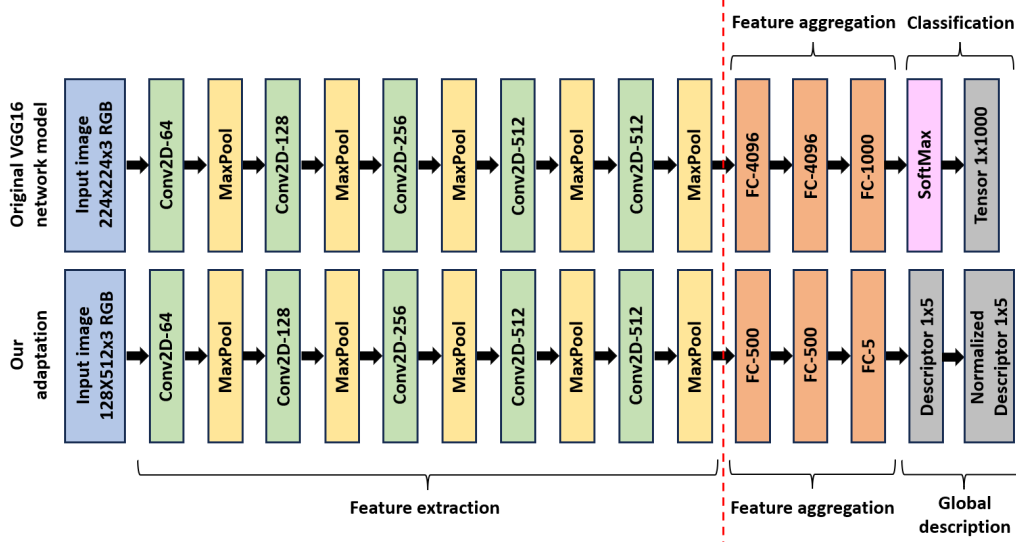


Figure 1. Original VGG16 network model (above) and our adaptation (below). Convolutional and max pooling layers have been left intact, whereas the fully connected layers have been modified in order to adapt the architecture to the size of the input images and obtain a five-element global descriptor. ReLU layers have not been included so as to simplify this figure.

During the training, the loss function compares the output provided by the neural network with the required output. Later, the optimizer algorithm will modify the network weights according to the committed error to optimize the value of the loss function and achieve a more accurate prediction. Therefore, triplet losses minimize their value when the anchor and positive inputs are predicted as similar and the negative input is predicted as different to the other two inputs. During the network training process, the chosen loss function is expected to have an important influence on the performance of the trained network. In this paper, an exhaustive study is conducted to assess the influence of the loss function in the accuracy of the network when it is trained to solve the localization problem.

- **Triplet Margin Loss (Triplet Loss):** This is the most renowned triplet loss. It returns the average value of all the batch combinations:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N [D_{a,p}^i - D_{a,n}^i + m]_+$$

where $D_{a,p}^i$ is the euclidean distance between the anchor and positive descriptors in the i -th triplet, $D_{a,n}^i$ is the euclidean distance between the anchor and negative descriptors, $[\dots]_+$ is the ReLU function, m is the margin and N is the batch size (number of triplet samples that are taken into account before updating the internal model parameters).

- **Lifted Embedding Loss:** This loss, described in [14], is characterized by not only taking into account the distance between the anchor and positive inputs and the distance between the anchor and negative inputs, but also trying to maximize the distance between the positive and negative inputs:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \left[D_{a,p}^i + \ln \left(e^{m-D_{a,n}^i} + e^{m-D_{p,n}^i} \right) \right]_+$$

where $D_{p,n}^i$ is the euclidean distance between the positive and negative descriptors in the i -th triplet sample.

- **Lazy Triplet Loss:** This loss returns the hardest example of the batch for the network learning process:

$$\mathcal{L} = \left[\max \left(\vec{D}_{a,p} - \vec{D}_{a,n} + m \right) \right]_+$$

where $\vec{D}_{a,p} = (D_{a,p}^1, D_{a,p}^2, \dots, D_{a,p}^N)$ are the euclidean distances between each anchor-positive pair and $\vec{D}_{a,n} = (D_{a,n}^1, D_{a,n}^2, \dots, D_{a,n}^N)$ are the euclidean distances between each anchor-negative pair.

- **Semi Hard Loss:** This loss is a Lazy Triplet Loss variant. It calculates the average distance between the anchor and positive descriptors, and the minimum distance between the anchor and negative descriptors. In other words, it returns the hardest negative example of the batch:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \left[D_{a,p}^i - \min \left(\vec{D}_{a,n} \right) + m \right]_+$$

- **Batch Hard Loss:** This loss is another variant of the Lazy Triplet Loss. It returns the maximum distance between the anchor and positive descriptors, and the minimum distance between the anchor and negative descriptors. Therefore, it returns the hardest positive and negative examples of the batch:

$$\mathcal{L} = \left[\max \left(\vec{D}_{a,p} \right) - \min \left(\vec{D}_{a,n} \right) + m \right]_+$$

- **Circle Loss:** This loss, proposed in [39], makes use of the cosine similarity metric instead of the euclidean distance:

$$\mathcal{L} = \ln \left(1 + \sum_{j=1}^N e^{\gamma \alpha_n^j s_n^j} + \sum_{i=1}^N e^{-\gamma \alpha_p^i s_p^i} \right)$$

where,

$$\alpha_p^i = [O_p - s_p^i]_+; \alpha_n^j = [s_n^j - O_n]_+; O_p = 1 - m; O_n = m$$

where s_p^i is the cosine similarity between the anchor and positive descriptors, s_n^j is the cosine similarity between the anchor and negative descriptors and γ is a scale factor.

- **Angular Loss:** This loss, introduced in [42], seeks to minimize the angle formed by the vector that connects the anchor and the negative descriptors and the vector that connects the positive and the negative descriptors. Thus, it minimizes the distance between the anchor and positive inputs:

$$\mathcal{L} = \ln \left(1 + \sum_{i=1}^N e^{f_{a,p,n}^i} \right)$$

where,

$$f_{a,p,n}^i = 4 \tan^2 \alpha (x_a^i + x_p^i)^T x_n^i - 2 (1 + \tan^2 \alpha) (x_a^i)^T x_p^i$$

where x_a^i is the anchor descriptor of the i -th triplet sample, x_p^i is the positive descriptor of the i -th triplet sample, x_n^i is the negative descriptor of the i -th triplet sample and α is an angular margin.

4. Visual Localization

With the aim of addressing the localization problem, the present work makes use of omnidirectional images captured in indoor environments by a catadioptric system mounted on a mobile robot. Subsequently, RGB images are converted to panoramic format with 128x512x3 pixels and split into training, validation and test sets. Additionally, a visual model is generated with the images used during the training process. For every image, the coordinates of the capture points are known (ground truth), which allows us to conduct a supervised training. Afterwards, we conduct the training, validation and test of the triplet network architecture proposed in section 3. In every stage, a triplet architecture will be used to train the network, in such a way that the model is trained with combinations of three images I_a, I_p, I_n , where each of the networks that compose the architecture receives an input image and outputs a descriptor of that image. In order to perform the validation and test of the network, any of the networks that form the triplet architecture, as they are identical, will be used to embed each test image into a global-appearance descriptor $\vec{d}_{test} \in \mathbb{R}^{5 \times 1}$ that will be compared with the rest of the image descriptors that constitute the visual map, composed of the images used during the training process. These descriptors are normalized and then compared using euclidean distance or cosine similarity. The nearest neighbour among the images in the visual model will allow us to estimate the position of the robot when it captured the test image. The next subsections describe the two localization approaches: hierarchical localization and global localization.

4.1. Hierarchical localization

Hierarchical localization involves estimating the coordinates where the robot has captured an image in two steps. First, we carry out a coarse localization, in which the network identifies the room where the robot is. Second, a fine localization is performed, in which the network determines the robot coordinates in the room that has been retrieved in the first stage.

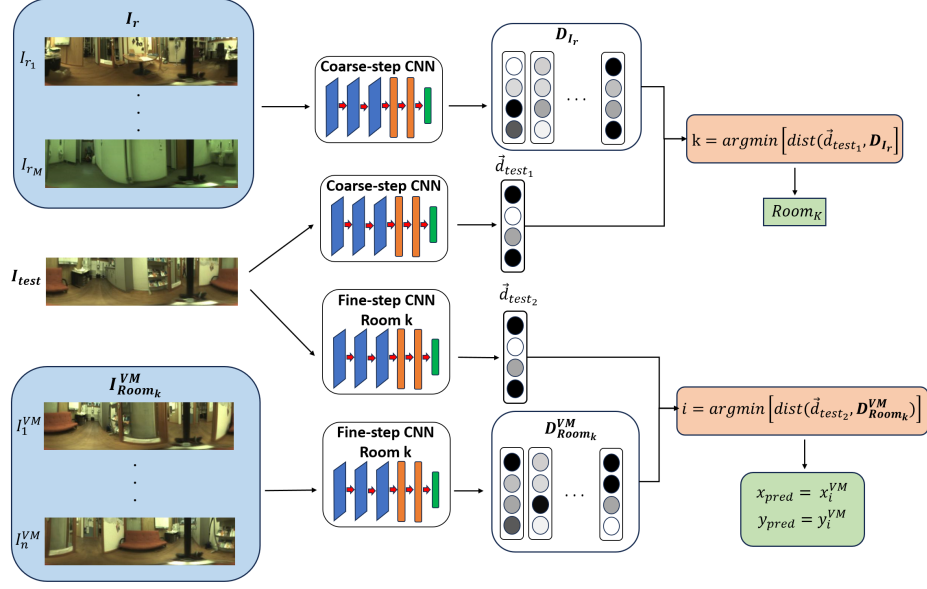


Figure 2. Hierarchical localization process performed in two steps. First, the test image descriptor \vec{d}_{test_1} is compared with representative descriptors of the rooms $D_{I_r} = [\vec{d}_{I_{r_1}}, \vec{d}_{I_{r_2}}, \dots, \vec{d}_{I_{r_M}}]$ and the nearest neighbour is considered the retrieved room k . Second, \vec{d}_{test_2} is compared with the descriptors of the images that compose the visual model of the retrieved room $D_{Room_k}^{VM} = [\vec{d}_1^{VM}, \vec{d}_2^{VM}, \dots, \vec{d}_n^{VM}]$ and the nearest neighbour indicates the robot coordinates inside the room (x_{pred}, y_{pred}) .

- Coarse localization:** in this stage, the network must determine in which room the test image has been taken. To do that, the triplet network is trained with combinations of three images I_a, I_p, I_n chosen randomly, in such a way that the anchor and positive images belong to the same room and the negative image must have been captured in a different room. The network is trained to output a descriptor per input image, with size 5×1 , as shown in Figure 1. Once trained, to test the network, the descriptor of each test image \vec{d}_{test_1} is compared with a set of descriptors that contain a representative descriptor of every room $D_{I_r} = [\vec{d}_{I_{r_1}}, \vec{d}_{I_{r_2}}, \dots, \vec{d}_{I_{r_M}}]$. The representative image of every room is the image captured from the position which is the closest to the geometrical centre of the room, where M is the number of rooms. If the predicted room matches the actual room, it will be considered as a network success.

- **Fine localization:** Once a room has been retrieved, the network must estimate the robot position inside the room. To do this part, an independent triplet network is trained for each one of the rooms, starting from the weights of the coarse-step network. In this case, all the training images belong to the same room and a distance threshold is defined to consider positive or negative pairs. In this work, the distance between anchor and positive images must be smaller than 0.3 m and the distance between anchor and negative images must be larger than 0.3 m. This threshold has not been chosen arbitrarily, since it is the minimum distance that permits every image to have at least one possible positive pair in the training dataset. To conduct the test, every test image descriptor \vec{d}_{test_2} is compared with the descriptor of every image that belongs to the visual model (VM) of the room that has been retrieved during the coarse localization $\mathbf{D}_{Room_k}^{VM} = [\vec{d}_1^{VM}, \vec{d}_2^{VM}, \dots, \vec{d}_n^{VM}]$, where n is the number of images in the visual model of the predicted room. The coordinates of the nearest neighbour are considered an estimation of the position of the robot when capturing the test image.

To address hierarchically the localization, one triplet network is trained to solve the coarse step, and one triplet network per room is trained to solve the fine localization step. In all cases, these networks are trained to provide a descriptor per input image. Once these networks are trained, the next steps are followed:

1. The robot captures an image I_{test} from an unknown position (x_{test}, y_{test}) .
2. The previously trained coarse-step network embeds the image into a global descriptor $\vec{d}_{test_1} \in \mathbb{R}^{5 \times 1}$.
3. The descriptor \vec{d}_{test_1} is compared with the representative descriptors $\mathbf{D}_{I_r} = [\vec{d}_{I_{r_1}}, \vec{d}_{I_{r_2}}, \dots, \vec{d}_{I_{r_M}}]$ via euclidean distance or cosine similarity. These descriptors are obtained from the representative image of each room using the coarse-step network.
4. The nearest neighbour indicates the retrieved room k .
5. The previously trained fine-step network of the retrieved room k embeds the image into a global descriptor $\vec{d}_{test_2} \in \mathbb{R}^{5 \times 1}$.
6. The descriptor \vec{d}_{test_2} is compared with the descriptors of the images that compose the visual model of the retrieved room $\mathbf{D}_{Room_k}^{VM} = [\vec{d}_1^{VM}, \vec{d}_2^{VM}, \dots, \vec{d}_n^{VM}]$.

These descriptors are obtained using the fine-step network of the retrieved room.

7. The coordinates of the nearest neighbour i are an estimation of the position of the robot inside the room $(x_{pred}, y_{pred}) = (x_i^{VM}, y_i^{VM})$ when capturing the test image.

4.2. Global localization

Global localization consists in determining the robot position in the entire map in one step. A unique network is trained for the whole environment, including images captured in all the rooms with random combinations. As in the fine localization, a distance threshold is set to create the positive and negative pairs: the distance between anchor and positive images must be smaller than 0.3 m and the distance between anchor and negative images must be larger than 0.3 m. In order to test the network, every test image descriptor \vec{d}_{test} is compared with the descriptors of the visual model of the whole map $\mathbf{D}^{VM} = [\vec{d}_1^{VM}, \vec{d}_2^{VM}, \dots, \vec{d}_n^{VM}]$, where n is the number of images in the complete visual model. Likewise, the coordinates of the nearest neighbour are considered an estimation of the position of the robot $(x_{pred}, y_{pred}) = (x_i^{VM}, y_i^{VM})$.

To address globally the localization, a unique triplet network is trained. This network is trained to provide a descriptor per input image. Afterwards, the next steps are followed:

1. The robot captures an image I_{test} from an unknown position (x_{test}, y_{test}) .
2. The trained network embeds the image into a global descriptor $\vec{d}_{test} \in \mathbb{R}^{5x1}$.
3. The descriptor \vec{d}_{test} is compared with the descriptors of the images that compose the visual model of the entire map $\mathbf{D}^{VM} = [\vec{d}_1^{VM}, \vec{d}_2^{VM}, \dots, \vec{d}_n^{VM}]$.
4. The coordinates of the nearest neighbour i are considered an estimation of the position of the robot $(x_{pred}, y_{pred}) = (x_i^{VM}, y_i^{VM})$ when capturing the test image.

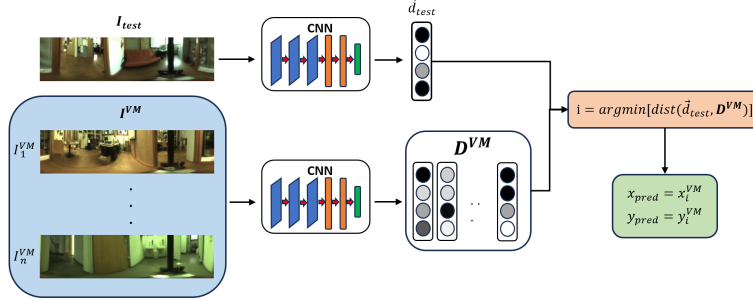


Figure 3. Global localization process performed in a unique step. Each test image descriptor \vec{d}_{test} is compared with the descriptors of the images that compose the visual model of the entire map $D^{VM} = [\vec{d}_1^{VM}, \vec{d}_2^{VM}, \dots, \vec{d}_n^{VM}]$ and the nearest neighbour indicates the robot coordinates (x_{pred}, y_{pred}) .

5. Experiments

This section describes the dataset and the results of the experimental evaluation. In this work, two experiments have been performed. Experiment 1 addresses a comparative evaluation of the influence of the triplet loss function in the performance of the network in a specific environment under different lighting conditions. Experiment 2 analyzes the performance of the network when different environments are considered at the same time.

5.1. Dataset

The images used in this work belong to COLD database (Pronobis and Caputo [32]). This dataset contains omnidirectional images captured by a mobile robot that makes use of a catadioptric vision system with a hyperbolic mirror. The robot follows a path inside several buildings and goes through different rooms, taking a picture every 0.08 s. Various types of rooms can be found inside the building, such as offices, a kitchen, a printer area, or a corridor that connects the different rooms. In this dataset, images captured under three illumination conditions can be found: Cloudy, Night and Sunny. Besides, some images include people moving or changes in the position of some pieces of furniture. All of this provides a complete dataset with plenty of defying examples due to illumination and dynamic changes. In this work, we have made use of three different environments: Freiburg (Part A, Path 2), Saarbrücken (Part A, Path 2) and Saarbrücken (Part B, Path 4). Despite the fact that two sets of images have been captured in the Saarbrücken building,

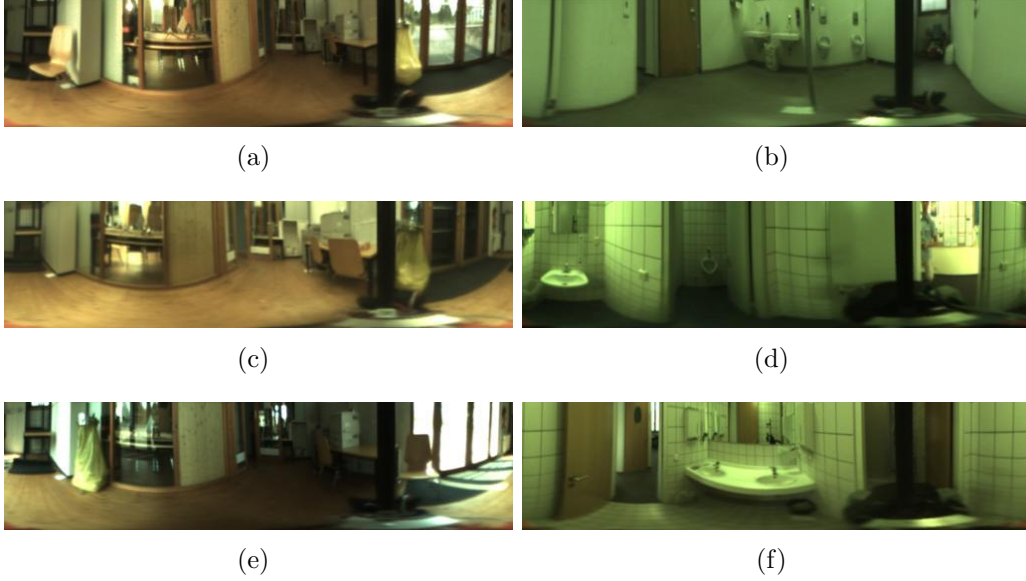


Figure 4. Examples of images captured under different lighting conditions (a) Cloudy, c) Night, e) Sunny) and examples of images captured in different environments (b) Freiburg, d) Saarbrücken A, f) Saarbrücken B).

they do not share any room, so they can be considered as two different environments.

Figure 4 shows some examples of images under each lighting condition and some examples of images that belong to each environment. These Figures illustrate some challenging cases that the network can find, such as changes of appearance caused by lighting variations or visual aliasing due to similar rooms that belong to different environments.

According to this philosophy, only cloudy images have been used to conduct the network training and validation, since it is the most standard illumination and it presents the lower contrast between the pixels corresponding to information indoors and outdoors, while all the illumination conditions are used for the test, so as to prove the network robustness against changes in the lighting conditions. Table 1 and Table 2 show the number of images from each image set used in Experiment 1 and Experiment 2, respectively. Table 3 enumerates the rooms in each environment and the number of images

| Image set | Illumination | Freiburg |
|---------------------|--------------|----------|
| Training | Cloudy | 588 |
| Validation 1 | Cloudy | 586 |
| Test 1 | Cloudy | 2595 |
| Test 2 | Night | 2707 |
| Test 3 | Sunny | 2114 |

Table 1. Size and lighting conditions of the training, validation and test sets used in Experiment 1.

| Image set | Illumination | Freiburg | Saarbrücken A | Saarbrücken B | TOTAL |
|---------------------|--------------|----------|---------------|---------------|-------|
| Training | Cloudy | 588 | 586 | 321 | 1495 |
| Validation 1 | Cloudy | 586 | 582 | 301 | 1469 |
| Validation 2 | Cloudy | 199 | 198 | 112 | 509 |
| Test 1 | Cloudy | 867 | 758 | 281 | 1906 |
| Test 2 | Night | 905 | 759 | 292 | 1956 |
| Test 3 | Sunny | 707 | X | 291 | 998 |

Table 2. Size and lighting conditions of the training, validation and test sets used in Experiment 2 (X indicates that the original Cold dataset contains no image in this set).

per room that shape the training set. The first row contains the id. of each room.

Additionally, the training set has also been employed as a visual map during the validation and the test. The training images have been captured roughly 20 cm apart from each other. The validation sets keep the same proportion of images per room as the training set. The test has been done for each illumination condition individually. Besides, training, validation and test sets do not share any of their images, that is, the validation and the test are carried out with images that the network has not seen during the training process. In Experiment 1, only Freiburg images have been used, whereas in Experiment 2 three different sets have been employed (Freiburg, Saarbrücken A and Saarbrücken B). While only one validation set has been used in Experiment 1, two different validation sets have been employed in

| Training set | 1PO | 2PO (1) | 2PO (2) | CNR | CR | KT | LO | PA | RL | ST | TL | TR | TOTAL |
|----------------------|-----|---------|---------|-----|-----|----|----|----|----|----|----|-----|-------|
| Freiburg | 44 | 47 | 40 | X | 235 | 45 | 40 | 57 | X | 40 | 40 | X | 588 |
| Saarbrücken A | 40 | 41 | X | 80 | 190 | X | X | 40 | 40 | X | 40 | 115 | 586 |
| Saarbrücken B | 40 | X | X | X | 129 | 44 | X | 40 | X | X | 68 | X | 321 |

Table 3. Number of images per room used during the training process.

Experiment 2. Validation 1 set has been used during the fine localization process to avoid having a too scarce number of validation images per room. Since that is not a problem in the rest of stages, Validation 2 set has been used to conduct the coarse localization and the global localization.

5.2. Experiment 1. Influence of the loss function.

In this experiment, a comparative evaluation has been done amongst different triplet loss functions (described in section 3). For all localization stages, a complete network training has been conducted with each triplet loss, giving different values to the parameters of the loss function with the purpose of finding their optimal value for each task. This experiment has been performed with the Freiburg set.

5.2.1. Hierarchical localization

a) Coarse localization

The aim of this stage is that the trained network is able to perform a room retrieval task. We have trained a network for each loss function and parameter setting, with a training length of 5 epochs and 50000 triplet samples per epoch. In the Table 4 the best results obtained with each loss function are shown.

| Loss function | Optimal parameters | Cloudy Accuracy(%) | Night Accuracy(%) | Sunny Accuracy(%) | Global Accuracy(%) |
|------------------|--------------------|--------------------|-------------------|-------------------|--------------------|
| Triplet Margin | m=1.25 | 99.23 | 97.04 | 95.08 | 97.12 |
| Lifted Embedding | m=0.25 | 99.11 | 97.12 | 93.52 | 96.58 |
| Circle | g=1,m=1 | 98.84 | 96.53 | 92.34 | 95.90 |
| Lazy Triplet | m=1.25 | 99.23 | 97.34 | 94.56 | 97.04 |
| Semi Hard | m=1 | 99.27 | 97.19 | 95.55 | 97.34 |
| Batch Hard | m=0.75 | 99.04 | 97.41 | 94.18 | 96.88 |
| Angular | a=30 ^o | 99.23 | 97.19 | 95.41 | 97.28 |

Table 4. Test accuracy for each loss function in the coarse localization.

This label shows that, in general terms, all the loss functions have led to a high accuracy in this stage. The best global accuracy is obtained with the Semi Hard Loss, which has output slightly better results than the Angular Loss and the Triplet Margin Loss. Moreover, the trained networks have been able to perform the localization task under every lighting condition, even

Sunny, which is the illumination that causes the largest change of appearance of the scenes compared with Cloudy, the condition used during the training process.

To verify the correctness of the results, confusion matrices will be presented, which show the network predictions, as well as its right and wrong predictions for each room. In the Figure 5 the confusion matrices of the network that provided the best results in the coarse localization are shown.

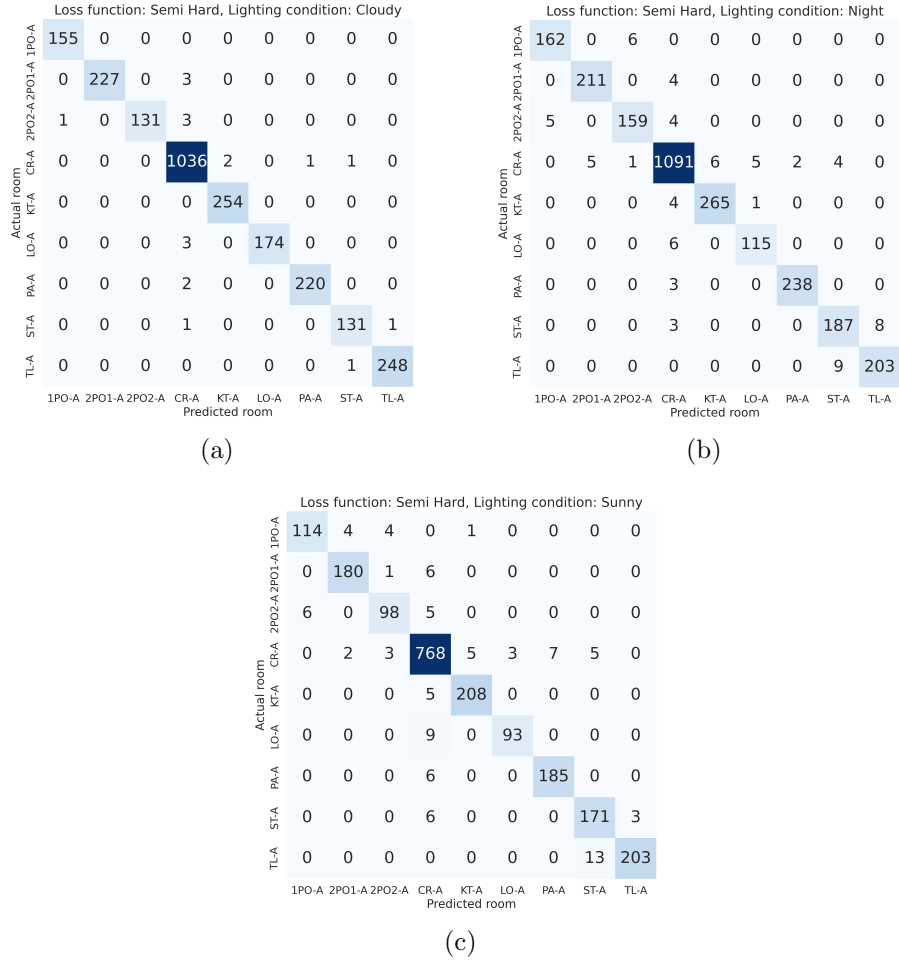


Figure 5. Confusion Matrices obtained in the test of the network trained with the Semi Hard Loss ($m=1$) for a) Cloudy, b) Night and c) Sunny.

These matrices indicate that most wrong predictions have occurred in the corridor (CR-A). That is logical, since it is the room with the largest dimensions and therefore much more images have been taken, and because the corridor is connected to five rooms. In fact, almost all wrong predictions happened between connected rooms as they share part of visual information in the transition areas. If only mistakes between non-connected rooms are considered, the network trained with the Semi Hard Loss would reach a 100% Cloudy accuracy, 99.96% Night accuracy and 99.72% Sunny accuracy. Besides, some of these mistakes are between resembling rooms like offices (1PO-A, 2PO1-A, 2PO2-A, LO-A). Another relevant number of errors take place between the stairs area (ST-A) and the toilet (TL-A), that can be caused by a small number of images from these rooms seen by the network during its training, which led to a network underfitting.

b) Fine localization

In this phase, a network is trained in order to estimate the robot position inside the room retrieved in the previous stage. For every room, a network has been trained with each loss function and its optimal parameters obtained in the coarse localization stage, with a training length of 5 epochs and 10000 triplet samples per epoch. Table 5 and Figure 6 reveal the average geometric error made by the network and the recall for the K nearest neighbours, respectively. The first row of table 5 also includes the minimum error that can be obtained in each experiment, given the distribution in the floor plane of the images in the visual model and the test images. Recall@K can be defined as the proportion of test images that the network can locate amongst the K nearest neighbours.

These results demonstrate that, in general terms, loss functions had a similar performance under Cloudy conditions. However, a larger difference can be appreciated under conditions that the network has not seen during the training process. The loss functions that provided the best results are the Semi Hard Loss and the Batch Hard Loss. In general terms, the errors are small for every lighting condition, especially Cloudy and Night. The errors obtained under Sunny conditions are larger because the mistakes committed during the coarse localization penalize the network performance in this stage.

Figure 7 shows the geometric error made in each room with the loss

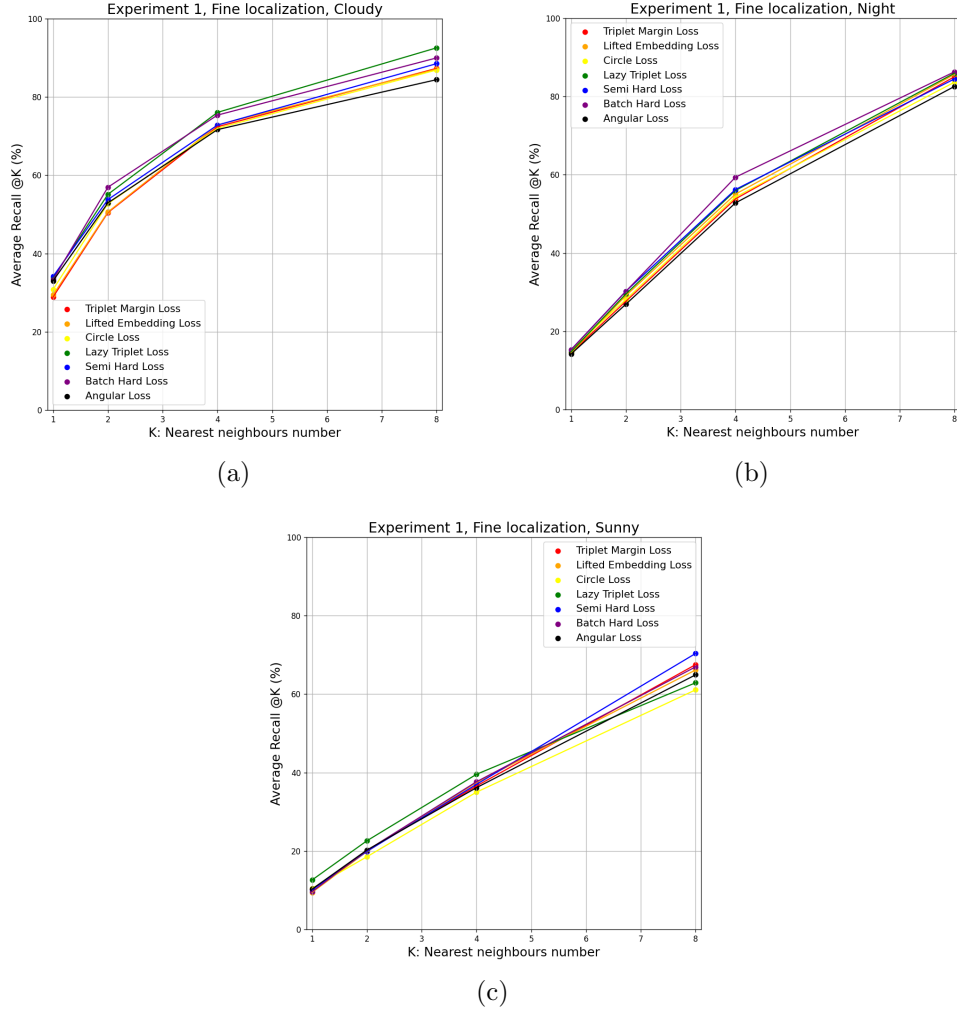


Figure 6. Average recall@K obtained in the fine localization with each loss function with a) Cloudy, b) Night and c) Sunny.

| Loss function | Cloudy Error (m) Min. error=0.128m | Night Error (m) Min. error=0.126m | Sunny Error (m) Min. error=0.120m | Global Error (m) |
|------------------|---------------------------------------|--------------------------------------|--------------------------------------|---------------------|
| Triplet Margin | 0.257 | 0.281 | 0.468 | 0.335 |
| Lifted Embedding | 0.252 | 0.310 | 0.667 | 0.410 |
| Circle | 0.292 | 0.373 | 0.746 | 0.470 |
| Lazy Triplet | 0.238 | 0.268 | 0.562 | 0.356 |
| Semi Hard | 0.249 | 0.275 | 0.398 | 0.307 |
| Batch Hard | 0.233 | 0.263 | 0.440 | 0.312 |
| Angular | 0.260 | 0.300 | 0.471 | 0.344 |

Table 5. Average geometric error (m) for each loss function in the fine localization.

function that provided the best results in the hierarchical localization and the minimum errors that can be reached (this is the piece of information that also appears in the first row of table 5). The error cannot be zero, because in order to happen that, the training and test sequences should be the same. The minimum reachable error is the one that would be obtained if the network had a 100% accuracy with $K=1$, in other words, if the predicted closest image always matches with the actual closest image. This is graphically shown in fig. 8.

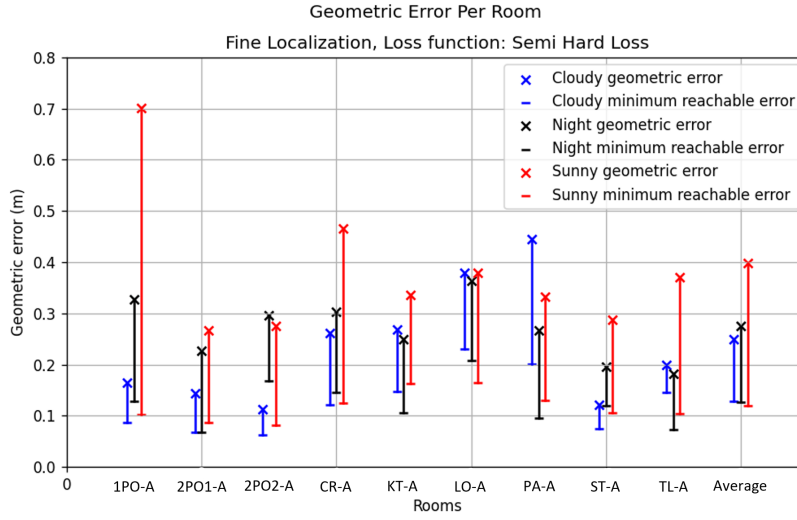


Figure 7. Geometric error (m) in every room with the Semi Hard Loss ($m=1$) in the fine localization.

From Figure 7 we notice that the error is very variable depending on the room, since an independent network has been trained for each room. The error depends as well on the difference between the paths followed by the robot in the training and test sequences. By comparing these results with the confusion matrices obtained from the coarse localization (Figure 5), we realize that the rooms where the geometric error is bigger are the rooms where more mistakes had been committed during the coarse localization. This can be observed more clearly in the room 1PO-A under Sunny conditions. The geometric error made in this room is 0.702 m, while the Sunny average error is 0.398 m.

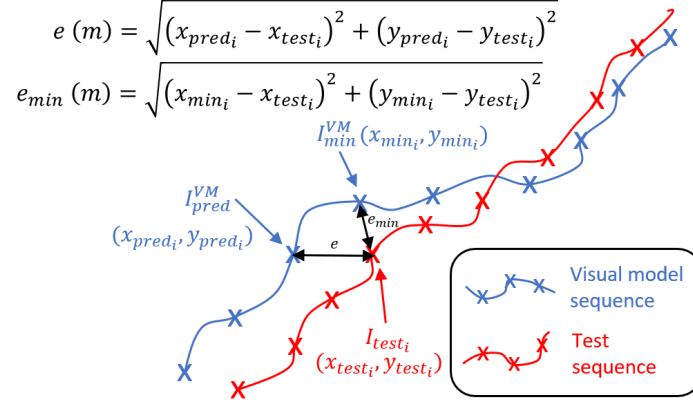
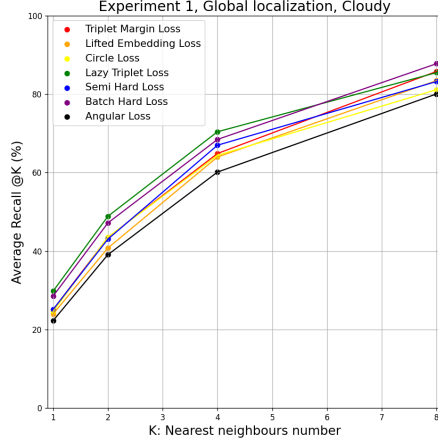


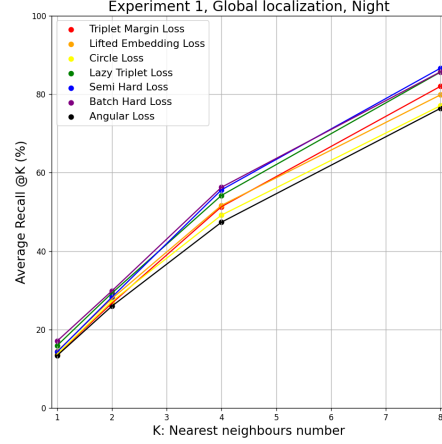
Figure 8. Error (m) and minimum reachable error (m) for the image I_{test_i} . Error $e(m)$ is the geometric distance between the capture point of the image I_{test_i} and the image of the visual model retrieved as the closest by the network I_{pred}^{VM} , whereas minimum error $e_{min}(m)$ is the geometric distance between the capture point of the image I_{test_i} and the actual closest image of the visual model I_{min}^{VM} .

5.2.2. Global localization

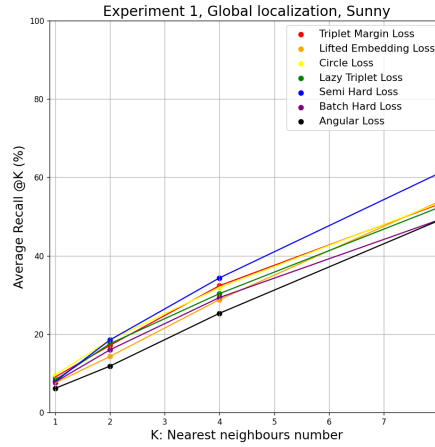
To address the global localization problem, an exhaustive study of the influence of the loss function and its parameters in the performance of the network has been done, as in the case of the coarse step of the hierarchical localization. A network has been trained for each loss function and parameters setting, with a training length of 5 epochs and 50000 triplet samples per epoch. Table 6 and Figure 9 reveal the average geometric error for each loss function and the recall for the K nearest neighbours, respectively.



(a)



(b)



(c)

Figure 9. Average recall obtained in the global localization with each loss function under a) Cloudy, b) Night and c) Sunny conditions.

Table 6 shows that the average errors tend to increase in the global localization in every lighting condition, comparing to the hierarchical localization. This is logical, since in this case the network tries to locate each image inside the entire map in a single step, and this environment is prone to visual aliasing, so the hierarchical process is able to better retain the features that characterize and distinguish every room. In general terms, comparing to

| Loss function | Optimal Parameters | Cloudy Error (m) Min. error=0.128m | Night Error (m) Min. error=0.126m | Sunny Error (m) Min. error=0.120m | Global Error (m) |
|------------------|--------------------|---------------------------------------|--------------------------------------|--------------------------------------|------------------|
| Triplet Margin | m=1 | 0.303 | 0.324 | 0.633 | 0.420 |
| Lifted Embedding | m=0.25 | 0.292 | 0.344 | 0.639 | 0.425 |
| Circle | $\gamma=1, m=1$ | 0.428 | 0.547 | 1.219 | 0.731 |
| Lazy Triplet | m=1.25 | 0.266 | 0.286 | 0.766 | 0.439 |
| Semi Hard | m=1.25 | 0.287 | 0.287 | 0.736 | 0.437 |
| Batch Hard | m=0.75 | 0.262 | 0.288 | 0.823 | 0.458 |
| Angular | $\alpha=30^\circ$ | 0.338 | 0.413 | 0.734 | 0.495 |

Table 6. Average geometric error (m) for each loss function in the global localization and optimal parameters.

hierarchical localization, the performance is slightly worse for Cloudy and Night, but the error is larger for Sunny. This can be observed clearly by comparing Figures 6 and 9. Under Cloudy and Night conditions, the network is able to locate a similar number of test images amongst the 8 closest neighbours. However, the recall is significantly lower in the global localization under Sunny. The loss functions that output the best results are the Triplet Margin Loss (m=1) and the Lifted Embedding Loss (m=0.25).

In the Figure 10, a comparison between the two proposed localization methods has been done by using maps with the predictions of the networks that had the best results: Semi Hard Loss, m=1, for the hierarchical localization and Triplet Margin Loss, m=1, for the global localization. The blue points represent the visual map, whilst the rest of points represent the test images. If the test image is located correctly amongst the K=1 nearest neighbours, the point will be green, if the image is located amongst K=2, the point will be light green, and so on until K=8 (please refer to the legends in Figure 10). If the image cannot be located amongst K=8, the point will be painted red. The lines connect every test image with the image of the visual map retrieved as the closest by the network. Moreover, Table 7 compares the average geometric error obtained with every loss function in the two proposed localization approaches: hierarchical and global.

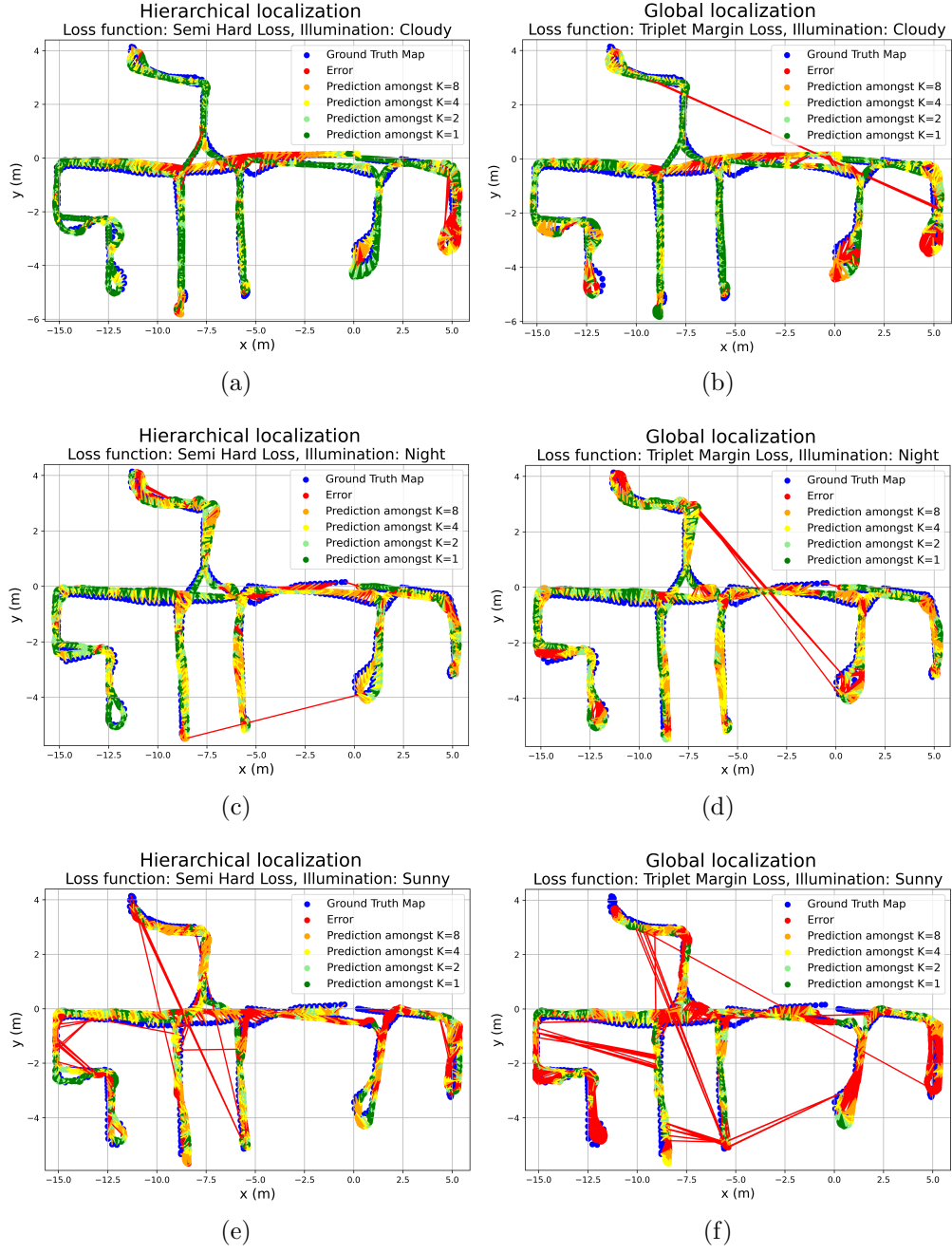


Figure 10. Network prediction for every test image in the hierarchical localization (a) Cloudy, c) Night, e) Sunny) and in the global localization (b) Cloudy, d) Night, f) Sunny).

| Loss function | Average Error (m) Hierarchical localization | Average Error (m) Global localization |
|------------------|--|--|
| Triplet Margin | 0.335 | 0.420 |
| Lifted Embedding | 0.410 | 0.425 |
| Circle | 0.470 | 0.731 |
| Lazy Triplet | 0.356 | 0.439 |
| Semi Hard | 0.307 | 0.437 |
| Batch Hard | 0.312 | 0.458 |
| Angular | 0.344 | 0.495 |

Table 7. Average geometric error (m) for each loss function in the hierarchical localization and in the global localization.

Table 7 shows that the hierarchical method permits performing a more accurate localization. For every loss function, the error made with the hierarchical localization is lower than the error made with the global localization. The maps represented in Figure 10 lead to the same conclusion. These maps also show very clearly that the number of errors between non-connected rooms is smaller in the hierarchical method, especially under Sunny conditions. In both methods, the errors take place more frequently in the transition zones or in junctions, as well as in zones where the robot turns, which means that the network is sensitive to changes in the orientation. Finally, we can observe that in both methods the networks have a better performance under Cloudy and Night conditions than Sunny. That is logical, since the network has been trained only with Cloudy images and Sunny is the lighting condition with the biggest change of appearance with respect to Cloudy. This could be fixed by training the network also with Sunny images, but the aim of this work is to prove the robustness of Triplet Networks against lighting changes and the results can be considered satisfactory.

Table 8 includes the localization time of each method. Localization time can be defined as the time gap since an image is captured until the coordinates of the image are obtained. In the conducted experiments, the hierarchical localization time is larger than global localization time. This difference is due to the fact that in the global localization, the image coordinates are retrieved in a single step by one network, whereas in the hierarchical localization two steps are needed. In this case, a single network is used during the coarse localization step and one network for every room is used during the fine localization step. However, in both cases the time is sufficiently low as to enable the robot to perform localization with a reasonable frequency.

| | Hierarchical localization | Global Localization |
|------------------------|---------------------------|---------------------|
| Localization time (ms) | 36.07 | 3.28 |

Table 8. Localization time (ms) for each localization method.

5.3. Experiment 2. Study of the performance of the network in different environments simultaneously.

In this experiment, the same procedure has been followed than in Experiment 1, with the difference that in this case, three image sets corresponding to different environments have been jointly used: Freiburg, Saarbrücken A and Saarbrücken B. Although the images in Saarbrücken A and Saarbrücken B have been captured in the same building, they do not share any rooms and therefore any visual information, so they must be considered as two different environments. Therefore, the networks to be trained are facing a more challenging task. The objective is to prove the ability of the networks in larger and different environments, and to explore the limits of the proposal. A network is retrained for each loss function with the optimal parameters obtained in Experiment 1.

5.3.1. Hierarchical localization

a) Coarse localization

In this stage, a network is trained for each loss function with combinations of three images I_a , I_p , I_n chosen randomly, in such a way that anchor and positive images belong to the same room and the negative image belongs to a different room. Therefore, in this case, the coarse step tries again to retrieve the room where the test image was captured, among the rooms present in the three environments to test. However, the negative image can belong to the same environment than the anchor or not. The training length for this part is 5 epochs and 50000 triplet samples per epoch.

We study first the ability of the network to retrieve the correct environment and second the ability to retrieve the correct room. In the Table 9 and Table 10 the best results obtained with each loss function are shown.

Table 9 proves that the trained network can retrieve the environment where an image has been captured in almost every case. Table 10 shows the

| Loss function | Cloudy Accuracy (%) | Night Accuracy (%) | Sunny Accuracy (%) | Global Accuracy(%) |
|------------------|------------------------|-----------------------|-----------------------|-----------------------|
| Triplet Margin | 99.90 | 99.69 | 99.70 | 99.76 |
| Lifted Embedding | 99.32 | 98.57 | 98.70 | 98.86 |
| Circle | 99.06 | 98.26 | 98.70 | 98.67 |
| Lazy Triplet | 99.74 | 99.23 | 99.90 | 99.62 |
| Semi Hard | 99.63 | 99.44 | 99.90 | 99.66 |
| Batch Hard | 99.90 | 97.96 | 100 | 99.28 |
| Angular | 97.64 | 95.50 | 98.30 | 97.15 |

Table 9. Environment retrieval accuracy with each loss function in the coarse localization.

| Loss function | Cloudy Accuracy (%) | Night Accuracy (%) | Sunny Accuracy (%) | Global Accuracy(%) |
|------------------|------------------------|-----------------------|-----------------------|-----------------------|
| Triplet Margin | 94.55 | 91.82 | 88.88 | 91.75 |
| Lifted Embedding | 96.75 | 93.61 | 92.59 | 94.31 |
| Circle | 94.28 | 89.52 | 88.38 | 90.73 |
| Lazy Triplet | 97.90 | 92.84 | 93.89 | 94.88 |
| Semi Hard | 97.85 | 95.30 | 91.48 | 94.88 |
| Batch Hard | 97.90 | 91.10 | 92.79 | 93.93 |
| Angular | 92.08 | 89.93 | 88.18 | 90.06 |

Table 10. Room retrieval accuracy with each loss function in the coarse localization.

accuracy of the network in the room retrieval task. As expected, the accuracy is lower than in Experiment 1, since now the network must distinguish amongst 22 rooms instead of 9. It should be noted that the accuracy obtained under Sunny conditions is higher than under Night because Sunny test set only contains images captured in two different environments (Freiburg and Saarbrücken B) and only 14 rooms are considered. This is due to the fact that the dataset does not contain any images captured under Sunny conditions in Saarbrücken A. In this stage, Semi Hard Loss and Lazy Triplet Loss have output the best results.

Figure 11 shows the confusion matrix obtained with the Semi Hard Loss in the environment retrieval task. This matrix reveals that the network is able to retrieve the environment as well as the room where an image has been captured, in general terms. The most wrong predictions between environments happened between Saarbrücken A and Saarbrücken B. Although the confusion matrix in the room retrieval task is not showed in this manuscript, the study of that matrix reveals leads to the conclusion that many of the network mistakes happened between similar rooms such as the corridors. If only

confusions between rooms of the same environment are considered, the errors mostly happen in the corridors, and some rooms such as the printer area (PA-A), the stairs (ST-A) or the toilet (TL-A) in Freiburg or the terminal room (TR-A) in Saarbrücken A contain a significant part of the errors.

Coarse localization, Environment retrieval
Loss function: Semi Hard Loss

| | | | | |
|--------------------|------|-----------------------|------|------|
| Actual environment | FR-A | 2475 | 4 | 0 |
| | SA-A | 1 | 1508 | 9 |
| | SA-B | 0 | 9 | 855 |
| | | FR-A | SA-A | SA-B |
| | | Predicted environment | | |

Figure 11. Confusion matrix obtained with the Semi Hard Loss in the environment retrieval. Cloudy, Night and Sunny tests are included in this matrix.

b) Fine localization

In this stage, a network per room is trained in order to determine the robot coordinates inside the room retrieved in the coarse localization. For every room, a network has been trained with each loss function, with a training length of 5 epochs and 10000 triplet samples per epoch. Table 11 and Figure 12 show the average geometric error made by the network and the recall for the K nearest neighbours, respectively. Network mistakes when retrieving the environment where an image has been captured have not been considered to calculate the average geometric error, because in this experiment, if the coarse-step network fails to retrieve the correct environment, no geometric error can be defined.

From these graphics we can observe that the error made by the network is larger than in Experiment 1. This is logical, since the network had a worse performance in the room retrieval as the number of rooms increased. In this case, the error committed under Cloudy conditions is substantially lower than under Night or Sunny. However, the errors are reasonable given

| Loss function | Cloudy Error (m) Min. error=0.136m | Night Error (m) Min. error=0.153m | Sunny Error (m) Min. error=0.108m | Global Error (m) |
|------------------|---------------------------------------|--------------------------------------|--------------------------------------|---------------------|
| Triplet Margin | 0.499 | 1.045 | 0.925 | 0.823 |
| Lifted Embedding | 0.399 | 0.869 | 0.490 | 0.586 |
| Circle | 0.623 | 1.327 | 1.081 | 1.010 |
| Lazy Triplet | 0.379 | 1.431 | 0.517 | 0.775 |
| Semi Hard | 0.379 | 0.848 | 0.504 | 0.577 |
| Batch Hard | 0.328 | 1.306 | 0.771 | 0.802 |
| Angular | 0.461 | 0.766 | 0.841 | 0.689 |

Table 11. Average geometric error made with each loss function in the fine localization.

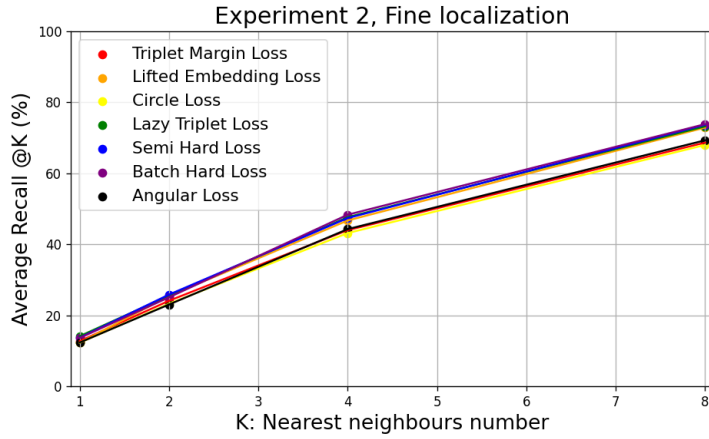


Figure 12. Average recall obtained in the hierarchical localization with each loss function.

the difficulty of the task. The loss functions that output the best results are the Semi Hard Loss and the Lifted Embedding Loss.

5.3.2. Global localization

In this part, a unique network is trained to localize the robot inside the entire map (containing the three environments considered in this section). A network has been trained for each loss function, with a training length of 5 epochs and 50000 triplet samples per epoch. Table 12 and Figure 13 show the average geometric error for each loss function and the recall for the K nearest neighbours, respectively. As in the fine localization, network mistakes when retrieving the environment where the robot is have not been considered to calculate the average geometric error.

| Loss function | Cloudy Error (m) Min. error=0.136m | Night Error (m) Min. error=0.153m | Sunny Error (m) Min. error=0.108m | Global Error (m) |
|------------------|---------------------------------------|--------------------------------------|--------------------------------------|---------------------|
| Triplet Margin | 0.570 | 1.010 | 0.970 | 0.850 |
| Lifted Embedding | 0.741 | 1.010 | 1.459 | 1.070 |
| Circle | 1.175 | 2.174 | 1.846 | 1.732 |
| Lazy Triplet | 0.696 | 0.903 | 1.235 | 0.945 |
| Semi Hard | 0.503 | 0.756 | 0.816 | 0.691 |
| Batch Hard | 0.453 | 0.830 | 0.812 | 0.698 |
| Angular | 0.786 | 1.390 | 1.703 | 1.293 |

Table 12. Average geometric error committed with each loss function in the global localization.

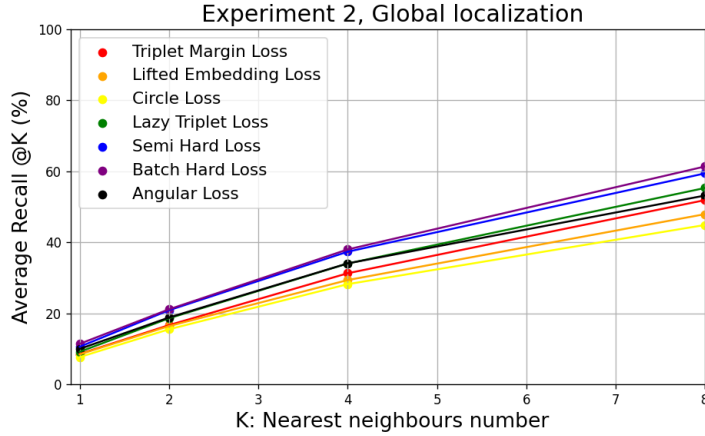


Figure 13. Average recall obtained in the global localization with each loss function.

As expected from the results of subsection 5.2.2, the performance of the network decreases in the global localization. In this case, the loss functions that output the best results are the Semi Hard Loss and the Batch Hard Loss.

Figure 14 shows the accuracy of the network in the environment retrieval task. In this stage, the network still retrieves the correct environment with high accuracy. Likewise, most errors take place between Saarbrücken A and Saarbrücken B.

If the two methods are compared, hierarchical localization enables us to do a more accurate localization. Table 13 reveals that every loss function

Global localization, Environment retrieval
Loss function: Semi Hard Loss

| | | | | |
|--------------------|------|-----------------------|------|------|
| Actual environment | FR-A | 2413 | 47 | 19 |
| | SA-A | 0 | 1411 | 107 |
| | SA-B | 0 | 84 | 780 |
| | | FR-A | SA-A | SA-B |
| | | Predicted environment | | |

Figure 14. Confusion matrix obtained with the Semi Hard Loss in the environment retrieval task in the global localization. Cloudy, Night and Sunny tests are included in this matrix.

| Loss function | Average Error (m) | |
|-----------------------|---------------------------|---------------------|
| | Hierarchical localization | Global localization |
| Triplet Margin Loss | 0.823 | 0.850 |
| Lifted Embedding Loss | 0.586 | 1.070 |
| Circle Loss | 1.010 | 1.732 |
| Lazy Triplet Loss | 0.777 | 0.945 |
| Semi Hard Loss | 0.577 | 0.691 |
| Batch Hard Loss | 0.802 | 0.698 |
| Angular Loss | 0.689 | 1.293 |

Table 13. Average geometric error (m) for each loss function in the hierarchical localization and in the global localization.

except the Batch Hard Loss presents a better performance in the hierarchical method.

5.4. Comparison with other works

Finally, the proposed method is compared with similar approaches that used global-appearance descriptors obtained with analytical techniques, such as gist or HOG ([31]), and with CNNs models that are adapted and retrained in order to tackle hierarchical localization in indoor environments. All the experiments have been conducted under similar conditions. All the approaches used a training set composed of images captured under cloudy conditions and tested their models under three different lighting conditions (cloudy, night and sunny). Table 14 shows the geometric error made in the hierarchical localization with each method. In the case of our method, the Semi Hard

| Global-Appearance Descriptor Technique | Cloudy Error (m) | Night Error (m) | Sunny Error (m) | Global Error (m) |
|---|---------------------|--------------------|--------------------|---------------------|
| SVM + K-NN [6] | 0.051 | 0.527 | 0.773 | 0.450 |
| Alexnet [4] | 0.293 | 0.288 | 0.690 | 0.424 |
| EfficientNet [35] | 0.240 | 0.330 | 0.440 | 0.337 |
| Triplet Network with Semi Hard (ours) | 0.249 | 0.275 | 0.398 | 0.307 |
| Triplet Network with Batch Hard (ours) | 0.233 | 0.263 | 0.440 | 0.312 |
| gist [5] | 0.052 | 1.065 | 0.884 | 0.667 |
| HOG [5] | 0.163 | 0.451 | 0.820 | 0.478 |

Table 14. Comparison with other methods in the complete hierarchical localization.

Loss and the Batch Hard Loss are considered, as they are the loss functions that have lead to the best global results.

Table 14 shows that the error made by our approach with the Batch Hard Loss under Cloudy conditions is slightly smaller to the error made by other approaches, which was already very small. In this case, some works such as [5] or [6] are not directly comparable to our method, since they used a denser visual map. Therefore, the error made in their method is lower than the minimum reachable error in our work (to see the minimum error that can be made with our method, please refer to the first row of Table 5).

Moreover, our approach outperforms the rest of works under lighting conditions that the networks have not seen during the training process, which proves that triplet architectures are accurate and robust tools to tackle visual localization problem in challenging environments without the need of a large dataset or a data augmentation.

6. Conclusions

Throughout the present work, two different localization approaches have been tackled (hierarchical and global) in indoor environments, with the use of triplet neural networks along with panoramic images. In the hierarchical method, a room retrieval task is performed in first place. Afterwards, the coordinates of the image inside the retrieved room in the first stage are determined. Meanwhile, in the global localization, the coordinates of the

image are estimated in a single step. The experiments demonstrate that the hierarchical approach performs a more accurate localization.

Triplet Networks are proposed to address both localization approaches. The VGG16 network model is adapted and retrained in such a way that the proposed architecture receives three different panoramic images and outputs three global appearance descriptors.

In order to test their robustness against significant changes of appearance, the networks have been trained with images captured under Cloudy conditions and tested under different lighting conditions. As expected, the trained networks tend to work better under Cloudy conditions, whereas the error made is larger under Night and especially Sunny.

Moreover, experiment 1 has addressed an exhaustive comparative evaluation of the influence of the triplet loss function in the performance of the network in every localization stage. In general terms, all the loss functions tend to output a high accuracy under Cloudy conditions. However, the results obtained with each loss function differ when the network is facing a more challenging task such as global localization or lighting conditions that the network has never seen during the training process. The loss function that showed the best performance is the Semi Hard Loss, and other losses such as the Batch Hard or the Triplet Margin also showed a good performance in difficult tasks.

Besides, experiment 2 considers three different environments simultaneously in order to evaluate the robustness of the proposed architecture in larger and repetitive environments and to explore the limits of the proposal. In this part, the accuracy is lower than in Experiment 1 as expected, since the trained network is facing a more challenging task. However, the errors made under every lighting condition are reasonably small given the difficulty of the problem. In this case, the Semi Hard Loss has also output the best results, which proves that it is a robust loss function to address different visual localization tasks.

Finally, our method has been compared with similar approaches that addressed a hierarchical localization. Under Cloudy conditions, our method has led to a similar error than other works, which was already small. However,

the error made by our approach is lower under Night and Sunny conditions (those that the networks have not seen during the training process). Therefore, triplet networks have proved to be a robust tool to address visual localization in challenging indoor environments.

In future works, the proposed architecture will be extended to outdoor environments, which are more challenging and show bigger changes of appearance. Furthermore, we will explore the use of more complex architectures to tackle visual localization in larger indoor environments.

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.

Data availability

The code used in this work is available in <https://github.com/MarcosAlfaro/TripletNetworksIndoorLocalization.git>.

Acknowledgments

This work is part of the project TED2021-130901B-I00 funded by MCIN/AEI/10.13039/501100011033 and by the European Union “NextGenerationEU” /PRTR, and of the project PROMETEO/2021/075 funded by Generalitat Valenciana.

References

- [1] Amorós, F., Payá, L., Mayol-Cuevas, W., Jiménez, L. M., & Reinoso, O. (2020). Holistic descriptors of omnidirectional color images and their performance in estimation of position and orientation. *IEEE Access*, 8, 81822-81848, <https://doi.org/10.1109/ACCESS.2020.2990996>.
- [2] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5297-5307), <https://doi.org/10.48550/arXiv.1511.07247>.

- [3] Ballesta, M., Paya, L., Cebollada, S., Reinoso, O., & Murcia, F. (2021). A cnn regression approach to mobile robot localization using omnidirectional images. *Applied Sciences*, 11(16), 7521, <https://doi.org/10.3390/app11167521>.
- [4] Cabrera, J. J., Cebollada, S., Flores, M., Reinoso, Ó., & Payá, L. (2022). Training, optimization and validation of a cnn for room retrieval and description of omnidirectional images. *SN Computer Science*, 3(4), 271, <https://doi.org/10.1007/s42979-022-01127-8>.
- [5] Cebollada, S., Payá, L., Jiang, X., & Reinoso, O. (2022). Development and use of a convolutional neural network for hierarchical appearance-based localization. *Artificial Intelligence Review*, 1-28, <https://doi.org/10.1007/s10462-021-10076-2>.
- [6] Cebollada, S., Payá, L., Peidró, A., Mayol, W., & Reinoso, O. (2023). Environment modeling and localization from datasets of omnidirectional scenes using machine learning techniques. *Neural Computing and Applications*, 1-22, <https://doi.org/10.1007/s00521-023-08515-y>.
- [7] Chen, X., Läbe, T., Milioto, A., Röhling, T., Behley, J., & Stachniss, C. (2022). OverlapNet: A siamese network for computing LiDAR scan similarity with applications to loop closing and localization. *Autonomous Robots*, 1-21, <https://doi.org/10.1007/s10514-021-09999-0>.
- [8] Chen, Y., Chen, R., Liu, M., Xiao, A., Wu, D., & Zhao, S. (2018). Indoor visual positioning aided by CNN-based image retrieval: training-free, 3D modeling-free. *Sensors*, 18(8), 2692, <https://doi.org/10.3390/s18082692>.
- [9] Cheng, D., Gong, Y., Zhou, S., Wang, J., & Zheng, N. (2016). Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1335-1344), <https://doi.org/10.1109/CVPR.2016.149>.
- [10] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). Ieee, <https://doi.org/10.1109/CVPR.2009.5206848>.

- [11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, <https://doi.org/10.48550/arXiv.2010.11929>.
- [12] Flores, M., Valiente, D., Gil, A., Reinoso, O., & Payá, L. (2022). Efficient probability-oriented feature matching using wide field-of-view imaging. Engineering Applications of Artificial Intelligence, 107, 104539, <https://doi.org/https://doi.org/10.1016/j.engappai.2021.104539>.
- [13] Foroughi, F., Chen, Z., & Wang, J. (2021). A cnn-based system for mobile robot navigation in indoor environments via visual localization with a small dataset. World Electric Vehicle Journal, 12(3), 134, <https://doi.org/10.3390/wevj12030134>.
- [14] Hermans, A., Beyer, L., & Leibe, B. (2017). In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737, <https://doi.org/10.48550/arXiv.1703.07737>.
- [15] Kallasi, F., Rizzini, D. L., & Caselli, S. (2016). Fast keypoint features from laser scanner for robot localization and mapping. IEEE Robotics and Automation Letters, 1(1), 176-183, <https://doi.org/10.1109/LRA.2016.2517210>.
- [16] Kim, S., Seo, M., Laptev, I., Cho, M., & Kwak, S. (2019). Deep metric learning beyond binary supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2288-2297), <https://doi.org/10.48550/arXiv.1904.09626>.
- [17] Kneip, L., Furgale, P., & Siegwart, R. (2013, May). Using multi-camera systems in robotics: Efficient solutions to the npnp problem. In 2013 IEEE International Conference on Robotics and Automation (pp. 3770-3776). IEEE, <https://doi.org/10.1109/ICRA.2013.6631107>.
- [18] Komorowski, J. (2021). Minkloc3d: Point cloud based large-scale place recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1790-1799), <https://doi.org/10.48550/arXiv.2011.04530>.

- [19] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84-90, <https://doi.org/10.1145/3065386>.
- [20] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324, <https://doi.org/10.1109/5.726791>.
- [21] Leyva-Vallina, M., Strisciuglio, N., & Petkov, N. (2021). Generalized contrastive optimization of siamese networks for place recognition. *arXiv preprint arXiv:2103.06638*, <https://doi.org/10.48550/arXiv.2103.06638>.
- [22] Li, D., Shi, X., Long, Q., Liu, S., Yang, W., Wang, F., ... & Qiao, F. (2020, October). DXSLAM: A robust and efficient visual SLAM system with deep features. In *2020 IEEE/RSJ International conference on intelligent robots and systems (IROS)* (pp. 4958-4965). IEEE, <https://doi.org/10.1109/IROS45743.2020.9340907>.
- [23] Lin, H. Y., Chung, Y. C., & Wang, M. L. (2021). Self-localization of mobile robots using a single catadioptric camera with line feature extraction. *Sensors*, 21(14), 4719, <https://doi.org/10.3390/s21144719>.
- [24] Liu, L., Li, H., & Dai, Y. (2019). Stochastic attraction-repulsion embedding for large scale image localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2570-2579), <https://doi.org/10.48550/arXiv.1808.08779>.
- [25] Liu, Y., & Huang, C. (2017). Scene classification via triplet networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(1), 220-237, <https://doi.org/10.1109/JSTARS.2017.2761800>.
- [26] Lopez-Antequera, M., Gomez-Ojeda, R., Petkov, N., & Gonzalez-Jimenez, J. (2017). Appearance-invariant place recognition by discriminatively training a convolutional neural network. *Pattern Recognition Letters*, 92, 89-95, <https://doi.org/10.1016/j.patrec.2017.04.017>.

- [27] Murillo, A. C., Guerrero, J. J., & Sagues, C. (2007, April). Surf features for efficient robot localization with omnidirectional images. In Proceedings 2007 IEEE International Conference on Robotics and Automation (pp. 3901-3907). IEEE, <https://doi.org/10.1109/ROBOT.2007.364077>.
- [28] Nilwong, S., Hossain, D., Kaneko, S. I., & Capi, G. (2019). Deep learning-based landmark detection for mobile robot outdoor localization. *Machines*, 7(2), 25, <https://doi.org/10.3390/machines7020025>.
- [29] Olid, D., Fácil, J. M., & Civera, J. (2018). Single-view place recognition under seasonal changes. arXiv preprint arXiv:1808.06516, <https://doi.org/10.48550/arXiv.1808.06516>.
- [30] Oliveira, G. L., Radwan, N., Burgard, W., & Brox, T. (2020). Topometric localization with deep learning. In *Robotics Research: The 18th International Symposium ISRR* (pp. 505-520). Springer International Publishing, <https://doi.org/10.48550/arXiv.1706.08775>.
- [31] Payá, L., Peidró, A., Amorós, F., Valiente, D., & Reinoso, O. (2018). Modeling environments hierarchically with omnidirectional imaging and global-appearance descriptors. *Remote sensing*, 10(4), 522, <https://doi.org/10.3390/rs10040522>.
- [32] Pronobis, A., & Caputo, B. (2009). COLD: The CoSy localization database. *The International Journal of Robotics Research*, 28(5), 588-594, <https://doi.org/10.1177/0278364909103912>.
- [33] Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 652-660), <https://doi.org/10.48550/arXiv.1612.00593>.
- [34] Qiu, K., Ai, Y., Tian, B., Wang, B., & Cao, D. (2018, June). Siamese-ResNet: Implementing loop closure detection based on Siamese network. In *2018 IEEE Intelligent Vehicles Symposium (IV)* (pp. 716-721). IEEE, <https://doi.org/10.1109/IVS.2018.8500465>.
- [35] Rostkowska, M., & Skrzypczyński, P. (2023). Optimizing Appearance-Based Localization with Catadioptric Cameras: Small-Footprint Models

- for Real-Time Inference on Edge Devices. *Sensors*, 23(14), 6485, <https://doi.org/10.3390/s23146485>.
- [36] Se, S., Lowe, D. G., & Little, J. J. (2005). Vision-based global localization and mapping for mobile robots. *IEEE Transactions on robotics*, 21(3), 364-375, <https://doi.org/10.1109/TR0.2004.839228>.
 - [37] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, <https://doi.org/10.48550/arXiv.1409.1556>.
 - [38] Su, Z., Zhou, X., Cheng, T., Zhang, H., Xu, B., & Chen, W. (2017, December). Global localization of a mobile robot using lidar and visual features. In *2017 IEEE international conference on robotics and biomimetics (ROBIO)* (pp. 2377-2383). IEEE, <https://doi.org/10.1109/ROBIO.2017.8324775>.
 - [39] Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., & Wei, Y. (2020). Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6398-6407), <https://doi.org/10.48550/arXiv.2002.10857>.
 - [40] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9), <https://doi.org/10.48550/arXiv.1409.4842>.
 - [41] Uy, M. A., & Lee, G. H. (2018). Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4470-4479), <https://doi.org/10.48550/arXiv.1804.03492>.
 - [42] Wang, J., Zhou, F., Wen, S., Liu, X., & Lin, Y. (2017). Deep metric learning with angular loss. In *Proceedings of the IEEE international conference on computer vision* (pp. 2593-2601), <https://doi.org/10.48550/arXiv.1708.01682>.
 - [43] Wozniak, P., Afrisal, H., Esparza, R. G., & Kwolek, B. (2018). Scene recognition for indoor localization of mobile robots using deep CNN.

- In Computer Vision and Graphics: International Conference, ICCVG 2018, Warsaw, Poland, September 17-19, 2018, Proceedings (pp. 137-147). Springer International Publishing, https://doi.org/10.1007/978-3-030-00692-1_13.
- [44] Xiao, L., Wang, J., Qiu, X., Rong, Z., and Zou, X. (2019). Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment. *Robotics and Autonomous Systems*, 117, 1-16, <https://doi.org/10.1016/j.robot.2019.03.012>.
 - [45] Xu, S., Chou, W., & Dong, H. (2019). A robust indoor localization system integrating visual localization aided by CNN-based image retrieval with Monte Carlo localization. *Sensors*, 19(2), 249, <https://doi.org/10.3390/s19020249>.
 - [46] Yin, H., Wang, Y., Ding, X., Tang, L., Huang, S., & Xiong, R. (2019). 3d lidar-based global localization using siamese neural network. *IEEE Transactions on Intelligent Transportation Systems*, 21(4), 1380-1392, <https://doi.org/10.1109/TITS.2019.2905046>.
 - [47] Yu, J., Zhu, C., Zhang, J., Huang, Q., & Tao, D. (2019). Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition. *IEEE transactions on neural networks and learning systems*, 31(2), 661-674, <https://doi.org/10.1109/TNNLS.2019.2908982>.
 - [48] Zhang, X., Su, Y., & Zhu, X. (2017, September). Loop closure detection for visual SLAM systems using convolutional neural network. In 2017 23rd International Conference on Automation and Computing (ICAC) (pp. 1-6). IEEE, <https://doi.org/10.23919/ICAC.2017.8082072>.
 - [49] Zhang, Z., & Peng, H. (2019). Deeper and wider siamese networks for real-time visual tracking. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4591-4600), <https://doi.org/10.48550/arXiv.1901.01660>.