Evolving Systems
https://doi.org/10.1007/s12530-024-09604-6

ORIGINAL PAPER

An evaluation of CNN models and data augmentation techniques in hierarchical localization of mobile robots

Juan José Cabrera[1] · Orlando José Céspedes[1] · Sergio Cebollada[1] · Oscar Reinoso[1,2] · Luis Payá[1]

Received: 31 January 2024 / Accepted: 25 June 2024

Abstract

This work presents an evaluation of CNN models and data augmentation to carry out the hierarchical localization of a mobile robot by using omnidirectional images. In this sense, an ablation study of different state-of-the-art CNN models used as backbone is presented and a variety of data augmentation visual effects are proposed for addressing the visual localization of the robot. The proposed method is based on the adaption and re-training of a CNN with a dual purpose: (1) to perform a rough localization step in which the model is used to predict the room from which an image was captured, and (2) to address the fine localization step, which consists in retrieving the most similar image of the visual map among those contained in the previously predicted room by means of a pairwise comparison between descriptors obtained from an intermediate layer of the CNN. In this sense, we evaluate the impact of different state-of-the-art CNN models such as ConvNeXt for addressing the proposed localization. Finally, a variety of data augmentation visual effects are separately employed for training the model and their impact is assessed. The performance of the resulting CNNs is evaluated under real operation conditions, including changes in the lighting conditions. Our code is publicly available on the project website https://github.com/juanjo-cabrera/IndoorLocalizationSingleCNN.git.

Keywords Mobile robotics · Omnidirectional vision · Hierarchical localization · Deep learning · Data augmentation

1 Introduction

In the ever-evolving landscape of Artificial Intelligence (AI), Convolutional Neural Networks (CNNs) have become a fundamental pillar of the technology, with disruptive problem-solving capabilities. This kind of neural networks were originally conceived for image recognition tasks, but have quickly transcended their initial boundaries, establishing themselves as a versatile and powerful tool for tackling a wide range of challenges in a variety of domains (LeCun and Bengio 1995).

The increasing use of CNNs can be attributed to their high ability to recognise patterns from different sources of information. This ability has made them essential in a wide variety of applications, from image recognition (Krizhevsky et al. 2012; Simonyan and Zisserman 2014) and object detection (Redmon et al. 2016; Ren et al. 2015) to semantic segmentation (Ronneberger et al. 2015) and even natural language processing (Kim 2014). The success of such architectures is based on their ability to extract features from data, which allows solving high-level problems such as visual localization.

In this sense, some researchers have addressed visual localization by means of 360° vision sensors due to its relatively low cost and the wide range of information they provide. When capturing images in real-world scenarios, especially in robotics applications, the environmental conditions can vary significantly. Consequently, addressing the visual localization could be particularly challenging due to

✉ Juan José Cabrera
juan.cabreram@umh.es

Orlando José Céspedes
orlando.cespedes@goumh.umh.es

Sergio Cebollada
s.cebollada@umh.es

Oscar Reinoso
o.reinoso@umh.es

Luis Payá
lpaya@umh.es

[1] Institute for Engineering Research (I3E), Miguel Hernandez University, Elche, Spain

[2] Valencian Graduate School and Research Network for Artificial Intelligence (valgrAI), Valencia, Spain

Published online: 08 July 2024

different phenomena such as changes in illumination conditions. For this reason, understanding and addressing the effects of illumination changes are crucial for developing robust CNN models.

Related with the above information, the main objective of this work is to analyze the influence of different visual effects applied to the training data in order to carry out the mapping and localization of a mobile robot, which moves in an indoor environment under real operation conditions. For this purpose, the omnidirectional images captured by a catadioptric vision sensor are used to train a CNN. Both the raw images, and some sets of images obtained after introducing visual effects to the original images in a data augmentation process are considered during the training. In this paper, we have also evaluated the performance of state-of-the-art CNN models when addressing localization through a hierarchical approach. In this sense, the CNN will be adapted and re-trained with a dual purpose: (1) to perform a rough localization step in which the model is used to predict the room from which a test image was captured, and (2) to address the fine localization step, which consists in retrieving the most similar image of the visual map among those contained in the previously predicted room by means of a pairwise comparison between descriptors obtained from an intermediate layer of the CNN. The main contributions of this paper can be summarized as follows.

- A CNN is adapted and re-trained to predict the room from which the robot captured an omnidirectional image which is transformed into panoramic. This approach enhances robotic localization by initially performing room recognition.
- We use the re-trained CNN to embed panoramic images into holistic descriptors by extracting the activation of an intermediate layer. These descriptors are compared to the visual model of the retrieved room via nearest neighbour search, providing an efficient method for scene recognition and position retrieval.
- We conduct a thorough study of the individual influence of different data augmentation visual effects when training a model to perform hierarchical localization. This analysis contributes to improve the robustness of the model and its generalization ability in localization tasks.
- We evaluate the performance of different state-of-the-art CNN models that are used as the backbone for the proposed localization task. This comparative evaluation provides valuable insights for selecting the most suitable CNN architecture for real-world localization applications.

This work is an extension of the initial developments presented in Céspedes et al. (2023). In this previous work, we used a basic CNN model (Places, Zhou et al. 2014) to perform the rough localization. However, our present proposal addresses both rough and fine localization steps and studies more exhaustively different state-of-the-art models such as AlexNet (Krizhevsky et al. 2012), ResNet-152 (He et al. 2016), ResNeXt-101 64x4d (Xie et al. 2017), MobileNetV3 (Howard et al. 2019), EfficientNetV2 (Tan and Le 2021) and ConvNeXt Large (Liu et al. 2022). Also, an ablation study of a variety of data augmentation visual effects are carried out with the aim of analysing the performance of the proposed tools under real operation conditions.

The following sections are structured as follows. First, in Sect. 2 we present a review of the state of the art on visual place-recognition and localization by means of artificial intelligence techniques. Second, in Sect. 3 we describe the proposed hierarchical localization method, the different CNN architectures which are evaluated and the proposed data augmentation visual effects. After that, we present in Sect. 4 the dataset used and the experiments carried out to test and validate the proposed method. Finally, conclusions and future works are outlined in Sect. 5.

2 State of the art

Artificial intelligence (AI) techniques are commonly proposed to address computer vision and robotics problems. Recent works, such as Aguilar et al. (2017), propose a pedestrian detector for Unmanned Aerial Vehicles (UAVs) based on Haar-LBP features combined with Adaboost and cascade classifiers with Meanshift. Another example is Wang et al. (2018), which utilizes an autoencoder for the fusion and extraction of multiple visual features from different sensors with the aim of carrying out motion planning based on deep reinforcement learning.

CNNs have proven to be successful in many practical applications. Well-known architectures, such as GoogLeNet (Szegedy et al. 2015), AlexNet (Krizhevsky et al. 2012) and VGG16 (Simonyan and Zisserman 2014) have been used as starting points to address new computer vision tasks. Regarding place-recognition, CNN models were firstly proposed to address this problem in Chen et al. (2014), where a pre-trained model called Overfeat (Sermanet et al. 2013) is used to extract features from images. Sünderhauf et al. (2015) provided a thorough investigation on the performance of extracted features for place recognition. In fact, they found out that the features extracted from convolutional layers were more robust against different lighting conditions than those extracted from fully connected layers which outperformed towards viewpoint changes. Bai et al. (2018) propose the SeqCNNSLAM method, which consists in using the pre-trained AlexNet (Krizhevsky et al. 2012) to extract features and feed the SeqSLAM algorithm (Milford and Wyeth 2012). Also Naseer et al. (2015) proposed a similar approach, but using GoogleNet (Szegedy et al. 2015). Some of the works have not only used images as source of information, but also point clouds (Uy and Lee 2018) and both combined (Komorowski et al. 2021).

In the context of robot localization, Kopitkov and Indelman (2018) propose using CNN holistic descriptors to estimate the robot position by learning a generative viewpoint-dependent model of CNN features with a spatially-varying Gaussian distribution. Sarlin et al. (2019) carry out a hierarchical modeling using a CNN, which extracts local features and holistic descriptors for 6-DOF localization. In that paper, a coarse localization is solved by using global descriptors, while a fine localization is solved by matching local features. Recent works (Cebollada et al. 2022) have proposed hierarchical visual models for efficient localization. This method involves arranging visual information hierarchically in different layers so that localization can be solved in two main steps. The first step involves coarse localization to roughly determine the area where the robot is located, and the second step involves fine localization within this pre-selected area.

Regarding the training of CNNs, a large and varied dataset is essential. Since a lack of a large enough datasets is quite common, Data Augmentation (DA) can be used to increase the training instances to avoid overfitting. As for the DA for a mobile robot localization task, it is essential to apply visual effects that may occur in real operation conditions to make the model robust against those effects. Considering as many effects as possible would increase the effectiveness of the CNN, but this would imply more processing power and memory. Numerous researchers have leveraged the data augmentation technique as a valuable tool to enhance the efficacy of their models. For example, Ding et al. (2016) train a CNN with three distinct types of data augmentation operations. Their investigation aims to enhance the performance of Synthetic Aperture Radar target recognition by achieving invariance against pose variations. Similarly, Salamon and Bello (2017) present a CNN designed for environmental sound classification, accompanied by an audio data augmentation strategy. This augmentation approach is useful to mitigate the scarcity of data in this domain, contributing to improved model performance. Furthermore, Perez and Wang (2017) present a study about the effectiveness of data augmentation to solve the classification task. Shorten and Khoshgoftaar (2019) present a survey about the existing methods for data augmentation and related developments. Nonetheless, the previously proposed data augmentation methods do not exactly analyze the visual phenomena that can occur when the mobile robot moves through the target environment under real-operation conditions. Therefore, the present work performs a data augmentation analysis that focuses on a wide range of those specific visual effects.

In light of the above information, the aim of this work is to analyze the influence of some visual effects to carry out data augmentation for CNN training to address a hierarchical localization (Cebollada et al. 2022). Hence, the efficiency of each visual effect will be assessed through the ability of the CNN model to robustly estimate the position where the image was captured. In addition, this work focuses on evaluating the performance of different well-known CNN models for both the coarse and fine localization steps. The first one consists in estimating the room where the image was taken by means of a classification final layer. The second one is addressed by extracting a global descriptor from an intermediate layer of the CNN and used to retrieve the most similar image that conforms the visual map. To address the proposed evaluation, the unique source of information is the set of images obtained by an omnidirectional vision sensor installed on the mobile robot, which moves in an indoor environment under real operation conditions.

3 Methodology

3.1 Hierarchical localization approach

This study aims to tackle visual localization through a hierarchical methodology by means of deep learning. The proposed approach (Fig. 1) consists of two main steps: an initial stage for rough localization, which consist in identifying the room from which the test image has been captured, and a subsequent phase for fine localization where the position of the robot is obtained by a pairwise comparison between the test image and the visual model that conforms the preselected room.

The initial step of rough localization is performed using the output of a CNN. The output layer of that CNN is composed by R neurons, each one corresponding to a room (R is the number of rooms or relevant areas in the target environment). Then, a SoftMax activation function is applied and the room prediction is obtained. However, before training the CNN, a dataset of labelled images captured along the target environment is needed. In this case, each image is labelled with the corresponding room information. The CNN is then trained to address the room retrieval task. Once the CNN is appropriately trained for the room classification task, the coarse localization step is performed: a test image im_{test} is fed into the CNN and the output indicates the room c_i in which the image was captured.

Simultaneously, a holistic descriptor is extracted by flattening the activation map from the last convolutional layer. This descriptor $\mathbf{d}_{test}$ is compared with the descriptors $D_{c_i} = \{\mathbf{d}_{c_i,1}, \mathbf{d}_{c_i,2}, \dots, \mathbf{d}_{c_i,N_i}\}$ from the visual map of

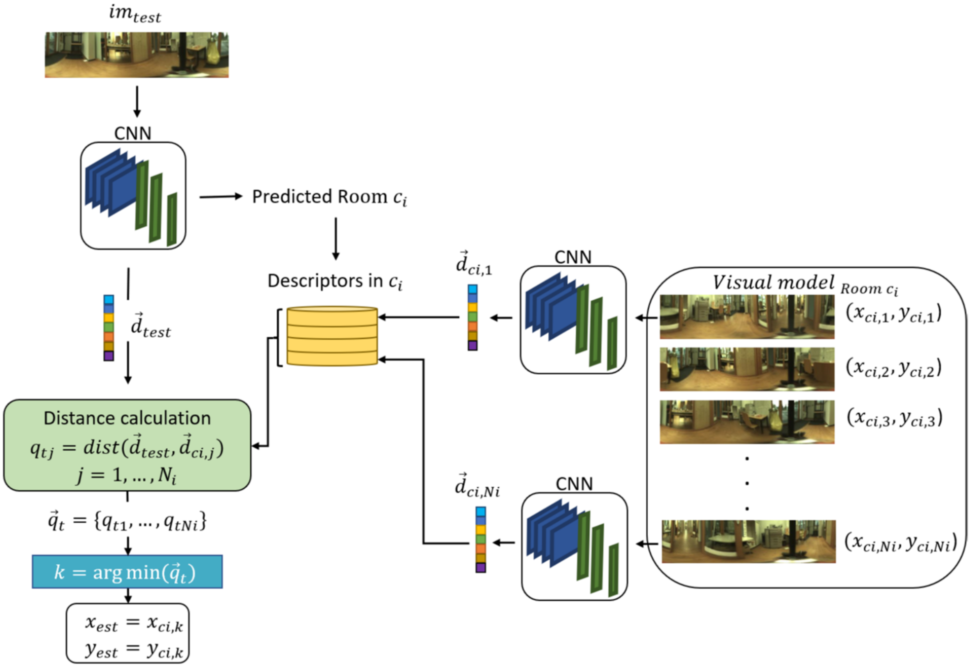

Fig. 1 Diagram of the proposed hierarchical localization. The test image im_{test} is the input of the CNN, which predicts the most likely room c_i and embeds the image into a global descriptor $\mathbf{d}_{test}$ by flattening the last activation map. This descriptor is compared with the descriptors from the training dataset included in the retrieved room by means of a nearest neighbour search. Consequently, the capture point of the image that corresponds to the most similar descriptor ($im_{c_i,k}$) is considered an estimation of the position where im_{test} was captured

the predicted room c_i, where N_i is the number of images in the room c_i. Note that the visual map descriptors are also obtained by flattening the last activation map of the same CNN. Then, the distance between the test descriptor $\mathbf{d}_{test}$ and each descriptor $\mathbf{d}_{c_i j} \in D_{c_i}$ in the room c_i is calculated (Eq. 1).

$$q_{t_j} = dist(\mathbf{d}_{test}, \mathbf{d}_{c_i j}), \quad j = 1, \dots, N_i \tag{1}$$

where N_i is the number of descriptors in room c_i and *dist* is the Euclidean distance (Eq. 2)

$$dist(\mathbf{d}_{test}, \mathbf{d}_{c_i j}) = \sqrt{\sum_{i=1}^{m} (d_{test,i} - d_{c_i j,i})^2} \tag{2}$$

where $\mathbf{d}_{test} = (d_{test,1}, d_{test,2}, \dots, d_{test,m})$ and $\mathbf{d}_{c_i j} = (d_{c_i j,1}, d_{c_i j,2}, \dots, d_{c_i j,m})$ are the descriptors of size m, and $d_{test,i}$ and $d_{c_i j,i}$ are the i-th components of the vectors $\mathbf{d}_{test}$ and $\mathbf{d}_{c_i j}$, respectively.

After that, a set $\mathbf{q}_t = \{q_{t1}, \dots, q_{tN_i}\}$ is constructed with the calculated distances. The index k which minimizes the distance in the set $\mathbf{q}_t$ is found in Eq. 3. Subsequently, the estimated position (x_{est}, y_{est}) corresponds to the position $(x_{c_i,k}, y_{c_i,k})$ from which the image $\mathbf{im}_{c_i,k}$ of the visual map (i.e, the image whose descriptor is the retrieved one $\mathbf{d}_{c_i,k}$) was captured (Eq. 4). This hierarchical approach ensures both a broad understanding of the scene and precise localization within the identified room, contributing to an effective visual localization strategy. Figure 1 outlines the whole localization process.

$$k = \arg\min(\mathbf{q}_t) \tag{3}$$

$$x_{est} = x_{c_i,k}, \quad y_{est} = y_{c_i,k} \tag{4}$$

3.2 CNN selection and adaption

Designing a Convolutional Neural Network to a address a specific task supposes a big challenge. In the present work, the CNN must be able to predict the room in which an image was captured and embed the input image into a global descriptor to retrieve the exact position within the predicted room. Crafting a CNN from scratch demands both a profound understanding of the specificities involved and access to a sufficiently varied dataset for effective training. Furthermore, as previously demonstrated in Ballesta et al. (2021), in general terms, re-training networks that have been designed for a different objective yields more precise and reliable outcomes in the new task than training from scratch.

In light of this information, this research work incorporates several widely recognised and tested CNN models, each of which serves as the backbone for our hierarchical localization task. These models cover a diverse range, addressing different architectural complexities and capabilities. All of the architectures employed were originally designed for visual object recognition. In this work, the CNN is first used to address the room retrieval problem, which is a similar task:

- AlexNet (Krizhevsky et al. 2012): AlexNet is a pioneering CNN architecture known for its success in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012. Comprising multiple convolutional and fully connected layers, AlexNet laid the foundation for subsequent CNN designs. This network and the following ones were trained to classify the 1.2 million high-resolution images into 1000 different classes. The weights and biases obtained by training with this database have been taken as starting point for our own task.

- ResNet-152 (He et al. 2016): ResNet, or Residual Network, introduced the concept of residual learning. This approach is based on skip connections and allows the CNN to learn an identity function. ResNet-152 is a specific variant featuring 152 layers, enabling the model to effectively capture intricate hierarchical features. Although it is computationally costly due to its depth, its accuracy and robustness compensate this cost.
- ResNeXt-101 64x4d (Xie et al. 2017): ResNeXt is an extension of the ResNet architecture, emphasizing a cardinality parameter to enhance model capacity. The cardinality is just the number of parallel blocks, that allows to learn various input representations. In this sense, ResNeXt-101 64x4d has a cardinality of 64. By increasing the cardinality, the network can capture a greater diversity of features, enhancing its potential ability to image recognition.
- MobileNetV3 (Howard et al. 2019): MobileNetV3 is designed for efficient mobile and edge computing applications. It uses depth-wise separable convolutions to build light weight deep neural networks. This fact makes them specially suitable for scenarios with resource constraints, such as performing the localization in real time by the robot's on-board computer.
- EfficientNetV2 (Tan and Le 2021): EfficientNetV2 is based on the EfficientNet architecture, and uses a technique called compound coefficient to scale up models in a simple but effective manner. It prioritizes model efficiency, achieving remarkable accuracy with fewer parameters compared to traditional CNNs. This makes EfficientNetV2 an attractive choice for applications requiring high accuracy with limited computational resources.
- ConvNeXt Large (Liu et al. 2022): ConvNeXt Large represents a recent advancement in CNN architectures. It leverages a combination of depth-wise separable convolutions, an inverted bottleneck and spatial factorization ("patchify"), contributing to improved efficiency and effectiveness in capturing features. Thus, outperforming the previous models in terms of accuracy.

By evaluating these diverse CNN models, we aim to comprehensively understand their strengths and weaknesses in the context of scene recognition and localization task. Regarding the room recognition, the final layer of all the architectures needs to be adapted for classifying the images into N categories corresponding to N possible rooms in the target environment ($N = 9$ in the dataset used in the present work, as described in Sect. 4.1). As for the fine localization, the global descriptor has been extracted by flattening the output of the Average Pooling Layer of each CNN model. Finally, Table 1 shows a summary with the evaluated models and its corresponding number of Floating Point Operations (FLOPs) and the number of parameters.

3.3 Data augmentation

Training a model involves setting up its parameters to perform a specific task. When a model has many parameters, it requires a sufficiently large number of examples for effective training. However, in practice, the training dataset is often limited. In such cases, data augmentation is a useful solution as it is able to generate new instances by applying various visual effects. This not only helps the model avoid overfitting but also makes it more robust against challenging real-operation dynamic conditions.

In previous studies focused on training models for visual localization, various effects like changes in orientation, reflections, alterations in illumination, noise, and occlusions were applied (Cabrera et al. 2022). The use of data augmentation has shown to improve model performance. These effects are applied individually or together to each image in the original dataset, and all the generated images are combined into a new augmented training dataset. However, the specific impact of each type of effect on the resulting CNN's performance is not well understood. This study aims to apply different data augmentation effects individually to evaluate their influence on the resulting CNN.

The focus of this work is on two categories of visual effects: changes in illumination conditions and changes in orientation. For changes in illumination conditions, the following effects are considered:

- Spotlights and shadows: Circular light sources, like bulbs, are common indoors. The proposed approach involves increasing pixel values to simulate higher light intensity (spotlights) and decreasing pixel values to sim-

Table 1 FLOPs and parameters of the evaluated and adapted models when the size of the input image is 512 × 128 × 3 pixels

Backbone model	FLOPs (G)	Number of parameters (M)
AlexNet	0.9	57.0
ResNet-152	15.2	58.2
ResNeXt-101 64X4d	20.4	81.4
MobileNetV3	0.3	4.2
EfficientNetV2	16.2	117.2
ConvNeXt Large	44.9	196.2

ulate shadows (shadow spots). Spotlights and shadow spots are applied separately for different data augmentation options. In our experiments, these bulbs are created with diameters ranging from 15 to 40 pixels. Five kinds of intensities variations are applied. In the first type the intensity is degraded ± 160 and in the fifth ± 100.

- General brightness and darkness: Low intensity values of the original images are increased to create brighter images, simulating higher overall illumination (e.g., a sunny day). Conversely, high intensity values are decreased to create darker images, simulating lower light supply (e.g., capturing images at night). Brightness and darkness are applied separately but used for the same data augmentation.
- Contrast: Image contrast plays a vital role in distinguishing objects in a scene. Images with low contrast tend to have a smoother appearance with fewer shadows and reflections. The contrast is modified following Eq. 5

$$I_s = 64 + c * (I - 64) \tag{5}$$

where I_s is the resulting image, I the original image and c is the contrast factor. For $c > 1$ the contrast increases and $c < 1$ decreases the contrast.

- Saturation: Color saturation, indicating the color intensity given by pixels, is considered. Lower saturation results in less colorful images, potentially resembling grayscale images for very low saturation. This phenomenon may occur in real environments and is incorporated into data augmentation. The color saturation can be adjusted by first converting the RGB image to HSV. Then, the saturation channel can be directly modified by multiplying it by a constant factor s. If the saturation is multiplied by $s > 1$, the colors become more saturated, whereas if multiplied by $s < 1$, the saturation decreases.

Regarding changes in orientation, these can occur during image capture when the robot captures images from the same position but with a different orientation. For this data augmentation option, new images are generated for each original image by applying rotations of n degrees, where $n = i \times 10°, i \in [1, 35]$. Thus, for each original image in the training set, 35 additional images are generated.

Figure 2 shows an example of the effects applied to a sample omnidirectional image converted to panoramic format. The first image corresponds to the original one and the rest of images include the different effects presented above (they have been separately applied).

4 Results

4.1 COLD Freiburg database

The current study utilizes images sourced from the Freiburg dataset, a subset of the COsy Localization Database (COLD) (Pronobis and Caputo 2009). This dataset contains omnidirectional images captured by a robot which follows various paths within a building at Freiburg University. The robot explores diverse spaces such as kitchens, corridors, printer areas, bathrooms, personal offices, and more. Image capture occurs under realistic operational conditions, including

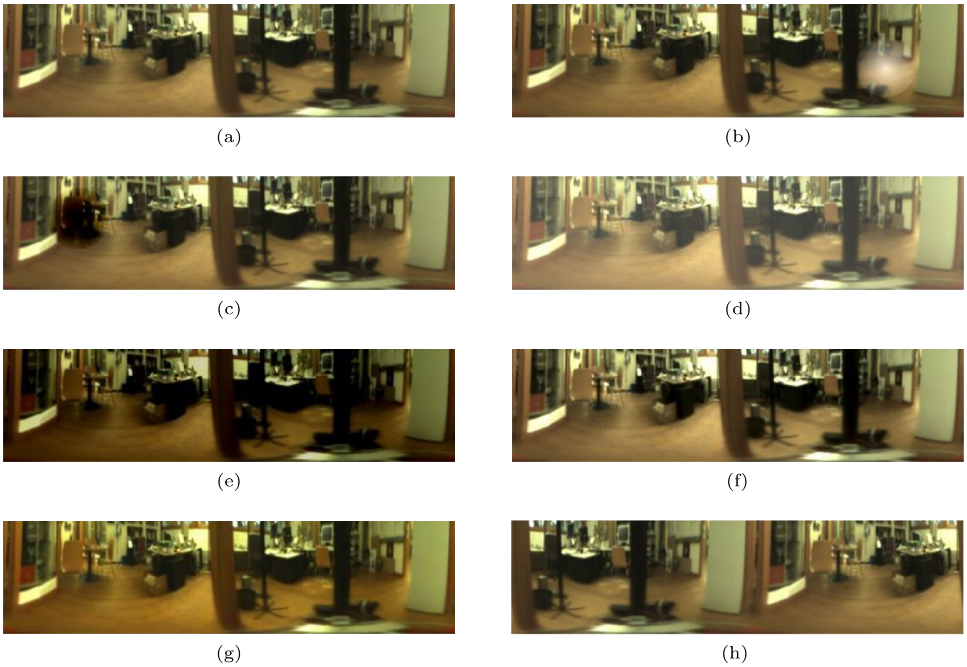
(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 2 Example of data augmentation where only one effect is applied over each image. **a** Original image, **b** spotlight effect, **c** shadow effect **d** general brightness, **e** general darkness, **f** contrast, **g** saturation and **h** rotation. The images contained in this dataset can be downloaded from the web site https://www.cas.kth.se/COLD/

changes in furniture arrangement, the dynamic presence of individuals in scenes, and fluctuations in illumination conditions, including cloudy days, sunny days, and nights.

To assess the impact of these variations on the localization task, we propose incorporating images taken exclusively on cloudy days as part of the training data. Additionally, a separate dataset comprising cloudy images (distinct from the aforementioned one) is employed as test set to evaluate localization performance without illumination changes. Furthermore, to appraise localization under varying illumination conditions, datasets captured on sunny days and at night are utilized as test sets. Beyond the images, the dataset offers ground truth data (obtained via a laser sensor), which is exclusively employed in this study to quantify localization errors. The ground truth over the path of the robot has been generated using the laser sensor in a grid-based SLAM technique, in particular, the one described in Grisetti et al. (2005, 2007). This solution, based on these two papers, can have an error up to 5 cm or 10 cm depending on the grid resolution.

Concerning the image capture process, the robot acquires images while it moves, introducing potential blur effects or dynamic alterations. Moreover, the chosen environment has the longest trajectory within the available database and is characterized by extensive windows and glass walls, making visual localization a particularly challenging problem. Consequently, this environment provides ideal conditions for evaluating the proposed localization methods under real operation conditions and real scenarios.

The selected dataset contains images from nine distinct rooms: a kitchen, a bathroom, a printer area, a stairwell, a long corridor and four offices. The cloudy dataset is downsampled to achieve an average distance of 20 cm between consecutive image capture points, resulting in the Baseline Training Dataset comprising 556 images. This dataset serves the dual purpose of training the CNNs and providing a visual map. In addition, a Validation Dataset is used during training and keeps the same proportion of images as the Baseline Training set. The Validation Dataset is also sampled at 20-cm intervals, but in this case in an interleaved manner with respect to the Baseline Training Dataset in such a way that the images in the baseline and validation datasets are different. In this regard, the validation covers uniformly the whole environment, which is expected to be a robust approach for validation, considering that the retrained CNN must be able to solve the localization problem considering the whole environment. Furthermore, the Baseline Training Dataset undergoes a data augmentation as described in Sect. 3.3, resulting in six additional training datasets. These datasets will be individually employed to train the CNNs, allowing an exploration of the impact of each visual effect on network performance. Table 2 shows a summary with the number of images per room of each training and validation dataset.

In terms of the test data, various datasets are considered: Cloudy Test Dataset, comprising images captured in cloudy conditions along a route distinct from training and validation sets (2595 images); Sunny Test Dataset, including all images captured in sunny conditions (2114 images); and Night Test Dataset, containing all images captured at night (2707 images). Table 3 shows a summary with the number of images per room of each test set. Consequently, network training and validation, in all instances, employs images captured exclusively in cloudy conditions, while testing occurs under three distinct lighting conditions: cloudy, sunny, or night. This methodology enables the assessment of the network's robustness against variations in lighting conditions.

Table 2 Number of images in each training dataset (number of images per room)

Training dataset	1P0-A	2P01-A	2P02-A	CR-A	KT-A	LO-A	PA-A	ST-A	TL-A
Baseline	44	46	31	238	46	26	57	30	38
Validation	43	47	32	236	46	26	57	31	38
Augmented 1	264	276	186	1428	276	156	342	180	228
Augmented 2	264	276	186	1428	276	156	342	180	228
Augmented 3	308	322	217	1666	322	182	399	210	266
Augmented 4	264	276	186	1428	276	156	342	180	228
Augmented 5	264	276	186	1428	276	156	342	180	228
Augmented 6	1364	1426	961	7378	1426	806	1767	930	1178

Table 3 Number of images in each test dataset (number of images per room)

Test dataset	1P0-A	2P01-A	2P02-A	CR-A	KT-A	LO-A	PA-A	ST-A	TL-A
Cloudy	155	230	135	1040	254	177	222	133	249
Night	168	215	168	1114	270	121	241	198	212
Sunny	123	187	109	793	213	102	191	180	216

4.2 Implementation details

In this work, the CNNs are trained to address the coarse localization or room retrieval stage. As this is a classification task, these networks have been retrained employing a Cross Entropy loss function (Eq. 6).

$$\mathcal{L}(y, \hat{y}) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{R} y_{ij}\log(\hat{y}_{ij}) \tag{6}$$

where y is the matrix of actual labels and $\hat{y}$ is the matrix of model predictions, both matrices have size $B \times R$, in which B is the number of samples (batch size) and R is the number of classes (rooms), y_{ij} is 1 if sample i belongs to class j and 0 otherwise, and $\hat{y}_{ij}$ is the probability predicted by the model that sample i belongs to class j.

In addition, Stochastic Gradient Descent (SGD) with Momentum 0.9 and Learning Rate of 0.001 has been used as optimization algorithm. Furthermore, the training batch size (B) was 16 and the total number of epochs was 30. For every architecture, the network that presents the best validation accuracy for room retrieval during the training is preserved for testing. Table 4 summarizes all the values of the parameters that have been described above.

All experiments are carried out with a NVIDIA GeForce RTX 3090 GPU with 24 GB. Our code is publicly available on the project website https://github.com/juanjo-cabrera/IndoorLocalizationSingleCNN.git.

4.3 CNN backbone ablation study

In this section, we asses an experimental evaluation of the different CNN models used as backbone presented in Sect. 3.2 for both rough and fine localization. As previously stated, the hierarchical localization proposed in this study comprises two distinct steps. The initial stage, rough localization step, involves retraining a model to execute the room retrieval task. Subsequently, the fine localization step utilizes the previously trained CNN to generate holistic descriptors, employing a nearest neighbor search method to estimate the precise position where an image was captured.

Table 4 Training parameters for room retrieval

Parameter	Value
Batch size (B)	16
Number of epochs	30
Learning rate	1×10^{-3}
Momentum	0.9
Number of rooms (R)	9

4.3.1 Coarse localization: room retrieval

This section presents the results derived from the use of different CNNs for the execution of the coarse localization or room retrieval stage. As described in Sect. 3.2, the CNN models evaluated in this article are AlexNet (Krizhevsky et al. 2012), ResNet-152 (He et al. 2016), ResNeXt-101 64x4d (Xie et al. 2017), MobileNetV3 (Howard et al. 2019), EfficientNetV2 (Tan and Le 2021) and ConvNeXt Large (Liu et al. 2022). The reason why we have selected these models is to cover a wide range of architectures proposed for image classification in the last 10 years.

The results in Table 5 showcase the performance of six different models used as backbone in the context of room retrieval across varied environmental conditions. In fact, each model was subjected to evaluation under cloudy, night, and sunny conditions, providing a comprehensive understanding of their robustness and adaptability to changes in environment illumination.

AlexNet exhibits an excellent overall performance, particularly in Cloudy conditions with an accuracy of 97.61%. In contrast, ResNet demonstrates robust performance but slightly lower accuracy compared to AlexNet. Notably, its accuracy decreases in sunny conditions which is the most demanding illumination environment. The ResNext model excels in cloudy conditions. However, it shows a comparatively lower accuracy in night scenarios. On the one hand, MobileNet stands out for its consistency, achieving high accuracy across all conditions. Its notable performance in sunny conditions, with an accuracy of 77.29%, highlights its generalisation capability. On the other hand, EfficientNet emerges as a top-performing model, outperforming others in terms of accuracy in cloudy and night scenarios, which are the most similar to training conditions. Finally, the most striking result comes from ConvNext, which consistently achieves the highest accuracy in all scenarios, making it the top-performing model. Particularly noteworthy is its

Table 5 Room retrieval ablation study for different top-level classification architectures tested under three different illumination conditions: cloudy, night, sunny and all together

Backbone model	Room retrieval accuracy (%)			
	Cloudy	Night	Sunny	Global
AlexNet	97.61	97.60	70.67	89.93
ResNet-152	96.76	96.64	64.95	87.63
ResNeXt-101 64X4d	98.11	95.16	72.47	89.71
MobileNetV3	98.50	96.93	77.29	91.88
EfficientNetV2	**98.81**	97.16	75.73	91.63
ConvNeXt Large	98.77	**97.64**	**86.28**	**94.80**

Bold values represent the best accuracy for every lighting condition exceptional accuracy of 86.28% in sunny conditions, indicating its robustness and generalization capabilities.

4.3.2 Fine localization

Once the CNN model is trained for the room retrieval step, it can be used to embed the input image into a global descripor. This facilitates the resolution of the fine localization step through an image retrieval process, in which the descriptor of the test image is compared with the descriptors of the visual map of the previously retrieved room. As in previous subsection, we are going to evaluate the performance of different CNN backbones to address the fine localization step. Fig. 3 shows the hierarchical localization error for different backbone models (AlexNet, ResNet-152, ResNeXt-101, MobileNetV3, EfficientNetV2 and ConvNeXt Large) under various lighting conditions (cloudy, night, sunny) and considering jointly the three conditions (global). The errors are measured in meters and are represented by box plots with whiskers, indicating the distribution of the errors. Furthermore, the Mean Absolute Error (Eq. 7) is represented by the black dot and the text displaying the error value. In addition, Table 6 shows the computation time required to execute the whole hierarchical localization process for all the evaluated models.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} \left|(x_i, y_i) - (\hat{x}_i, \hat{y}_i)\right| \tag{7}$$

where (x_i, y_i) is the actual position, $(\hat{x}_i, \hat{y}_i)$ is the position of the visual map retrieved after the complete localization process, and N is the number of images in the test dataset.

Each backbone model exhibited similar characteristics in hierarchical localization comparing to room retrieval, since both tasks are correlated. As Fig. 3 shows, AlexNet demonstrated a consistent localization error and low dispersion for cloudy and night conditions. However, its performance degraded in sunny conditions. ResNet-152 displayed higher errors across all conditions compared to AlexNet, with a notable increase of both the mean absolute error and dispersion in sunny conditions. ResNeXt-101 demonstrated a better performance than ResNet-152 for cloudy and sunny conditions, but the error slightly increases for night scenarios. MobileNet consistently maintained low errors across all conditions, signifying its adaptability to diverse lighting environments. EfficientNet showcased a worse performance than MobileNet in each scenario. Finally, ConvNeXt emerged as the top-performing model, consistently outperforming others with the lowest errors across all conditions. Its remarkable accuracy in sunny conditions implies a robust capability to handle scenarios with substantial changes of the lighting conditions. In terms of computation time, Table 6 illustrates that the hierarchical localization process with the shortest average computation time occurs when employing AlexNet, which requires only 3.4 ms. In contrast, the hierarchical localization process employing ConvNeXt Large requires the longest computation time, with a mean of 12.5 ms. However, despite the need for more time to estimate the

Table 6 Computation time required to execute the whole hierarchical localization process for all the evaluated models

Backbone model	Mean time (ms)
AlexNet	3.4
ResNet-152	6.9
ResNeXt-101 64X4d	9.5
MobileNetV3	4.6
EfficientNetV2	10.7
ConvNeXt Large	12.5

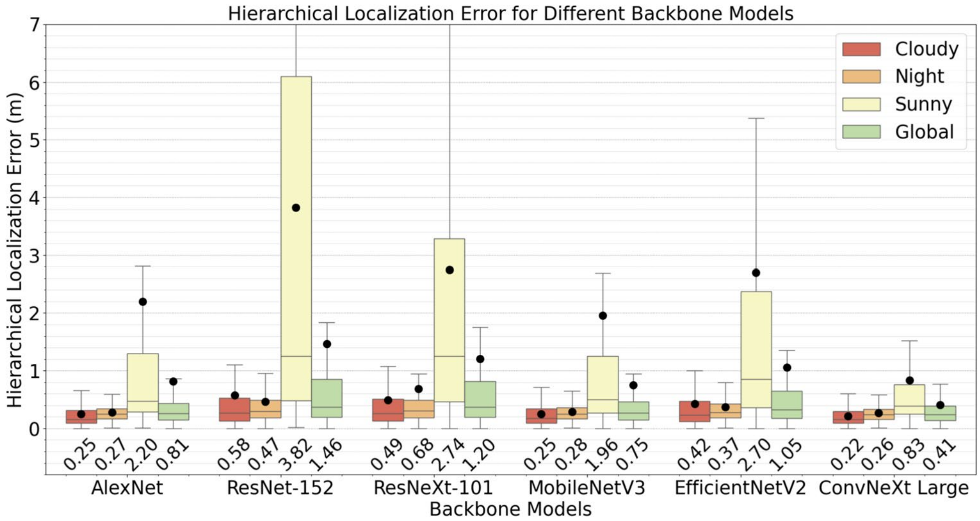

Fig. 3 Hierarchical localization errors in meters for different CNN architectures. The box plots represent the distribution of errors, with whiskers indicating variability. The Mean Absolute Error for each model and condition is marked by a black dot and annotated with the specific error value. Results are obtained under different lighting conditions: cloudy (red), night (orange), sunny (yellow) and considering jointly the three conditions (green)

position, this time is sufficiently short to enable real-time localization.

4.4 Data augmentation ablation study

In this comprehensive experiment, the investigation is extended to evaluate the influence of both data augmentation effects (illumination and orientation changes) on the performance of the CNN. Due to the existence of a high probability of variations in robot orientation during operation under real operation conditions with respect to the images captured in the visual map, a model should demonstrate robustness to orientation changes. To this end, a data augmentation technique is employed that consists in applying 35 different orientation changes to each training image as described in Sect. 3.3. This augmentation is essential to improve the adaptability of the model to the various orientations encountered in practice.

Simultaneously, the illumination effects that occur under real operating conditions, a critical aspect for robust visual perception, have been explored in detail. Five specific lighting effects are considered (Sect. 3.3): spotlights, shadow spots, general brightness/darkness, contrast, and saturation. Each effect is systematically applied individually on the training data set, leading to the creation of distinct augmented training datasets. Using the different effects separately allows a detailed understanding of their individual contributions, which sheds light on the importance of each effect in performance.

In particular, for each image, the experiment incorporates a detailed approach by applying different levels of spotlights, contrast and saturation (five levels for each), ensuring a thorough assessment of the impact of these factors on the ability of the CNN to adapt to various lighting conditions. In addition, the effect of brightness is meticulously explored, with three levels of brightness and three levels of darkness applied to each image. This dual investigation of orientation changes and illumination effects is intended to provide a comprehensive understanding of the robustness of the CNN to cope with real-world challenges, encompassing variations in both spatial orientation and illumination conditions. As a result of applying these effects, six additional training datasets have been obtained: Augmented Training Dataset 1 (spotlights), Augmented Training Dataset 2 (shadows), Augmented Training Dataset 3 (general brightness/darkness), Augmented Training Dataset 4 (contrast), Augmented Training Dataset 5 (saturation) and Augmented Training Dataset 6 (rotations). Augmented Training Datasets 1, 2, 4 and 5 consist of 3336 images each, whereas Augmented Training Datasets 3 and 6 includes 3892 and 17,236 images respectively.

In conclusion, in this ablation study the model is retrained using separately each of the Augmented Training Datasets 1, 2, 3, 4, 5 and 6. As in previous experiments, the Baseline Training Dataset serves as a visual map and the Validation Dataset is employed to validate the performance of the CNN. Furthermore, for the model evaluation, three different test datasets are considered: the Cloudy Test Dataset, the Night Test Dataset and the Sunny Test Dataset.

4.4.1 Coarse localization: room retrieval

In this subsection we use the best CNN architecture obtained in Sect. 4.3.1, which is ConvNeXt Large. In a similar approach, we have departed from the pre-trained weights for ImageNet Large Scale Visual Recognition Challenge and re-trained the model for the different datasets obtained by the proposed data augmentation.

Table 7 presents the room retrieval accuracy when the model has been trained with each of the augmented training datasets previously described. The performance of the CNN is evaluated under the three different lighting conditions: cloudy, night, sunny and all together.

Training with the baseline dataset shows a remarkable accuracy, especially in cloudy and night conditions. However, a significant decrease is observed in sunny conditions, which differ more from the training set. This evaluation provides a reference to analyse the impact of the different effects that have been applied to the training data.

The spotlight augmentation (Augmentation 1) shows insignificant improvements or even small decreases under night and sunny conditions. In contrast, data augmentation with shadows (Augmentation 2) produces slight improvements, especially in sunny conditions.

Alterations to the overall brightness and darkness of the image (Augmentation 3) are effective and show substantial improvements, especially in sunny conditions. In addition, contrast-based effects (Augmentation 4) are very effective, with substantial improvements in all lighting conditions and

Table 7 Room retrieval accuracy for ConvNeXt Large architecture with different augmented training datasets

Training dataset	Room retrieval accuracy (%)			
	Cloudy	Night	Sunny	Global
Baseline	98.77	**97.64**	86.28	94.80
Augmented 1 (spotlights)	98.84	97.45	86.14	94.71
Augmented 2 (shadows)	98.96	97.56	86.52	94.90
Augmented 3 (brightness/darkness)	98.81	97.41	91.11	96.10
Augmented 4 (contrast)	99.08	97.27	**93.57**	**96.84**
Augmented 5 (saturation)	98.88	97.60	83.07	93.91
Augmented 6 (rotations)	**99.15**	97.52	91.39	96.34

Bold values represent the best accuracy for every lighting condition especially in sunny circumstances, thus achieving improved results in this challenging environment.

Surprisingly, augmentation with changes in saturation (Augmented 5) shows a negative impact on accuracy, especially in sunny conditions. Finally, augmenting the data set with rotations (Augmented 6) shows substantial improvements, especially in cloudy conditions.

4.4.2 Fine localization

Once the ConvNeXt Large model is trained for the room retrieval step, it can be used to embed the input image into a global descriptor. This facilitates the resolution of the fine localization step through an image retrieval process, wherein the descriptor of the test image is compared with the descriptors of the visual map. As in previous subsection, we are going to evaluate the performance of different data augmentation effects to address the fine localization step.

As shown in Fig. 4, training with every augmented dataset result in similar network performance under cloudy illumination conditions for the fine localization task, achieving a mean absolute error around 0.22 ms. The same happens under the night condition, in which the mean absolute error is around 0.27 ms. In this case, the minimum error is obtained by training the network without data augmentation.

In contrast, under sunny lighting conditions the mean localization error has a higher variability, similarly to the coarse localization (Table 7). This demonstrates the correlation between the two tasks. Under this condition, the best fine localization result is obtained by training the model with the contrast effect (DA 4) and the worst with saturation (DA 5).

4.4.3 General comparison with other methods

Finally, the proposed method is compared with other previous global appearance techniques, including the use of single CNN structures (Cabrera et al. 2022; Rostkowska and Skrzypczynski 2023), triplet structures (Alfaro et al. 2024) and two classical analytical descriptors: HOG and gist, as described in Cebollada et al. (2022). Both HOG and gist are only taken into consideration when testing with night and sunny conditions, since the conditions of the cloudy test experiment in Cebollada et al. (2022) are different to the conditions in the present work. Table 8 compares all the methods in a global localization task, using in all cases the COLD-Freiburg dataset, which is the same dataset used in

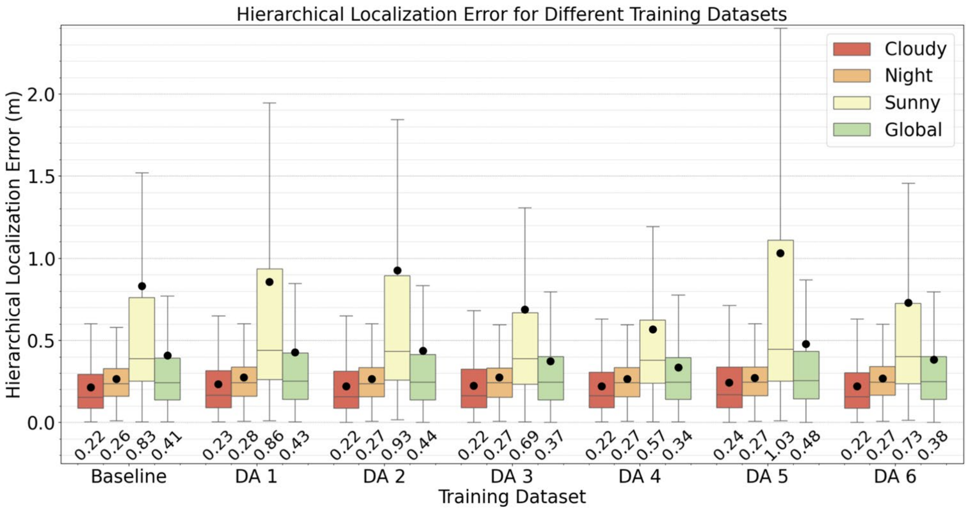

Fig. 4 Hierarchical localization errors in meters when training the ConvNeXt Large architecture with different data augmentation effects. The box plots represent the distribution of errors, with whiskers indicating variability. The Mean Absolute Error for each model and condition is marked by a black dot and annotated with the specific error value. Results are obtained under different lighting conditions: cloudy (red), night (orange), sunny (yellow) and considering jointly the three conditions (green)

Table 8 Comparison with other methods

Global-appearance descriptor technique	Cloudy error (m)	Night error (m)	Sunny error (m)
Alexnet + DA (Cabrera et al. 2022)	0.29	0.29	0.69
EfficientNet (Rostkowska and Skrzypczynski 2023)	0.24	0.33	0.44
Triplet VGG16 (Alfaro et al. 2024)	0.25	0.28	**0.40**
ConvNeXt Large (ours)	**0.22**	**0.26**	0.83
ConvNeXt Large + DA (ours)	**0.22**	0.27	0.57
HOG (Cebollada et al. 2022)	–	0.45	0.82
gist (Cebollada et al. 2022)	–	1.07	0.88

Bold values represent the minimum error for every lighting condition the previous subsections. This table shows that ConvNeXt Large without data augmentation provides the best results in terms of localization error for cloudy and night conditions. Training with data augmentation does not improve the performance in cloudy conditions. However, it favours the results under sunny conditions. In this illumination condition, the best result is obtained with a triplet VGG16 proposed in Alfaro et al. (2024).

5 Conclusion

This study assesses the application of a deep learning technique in addressing hierarchical localization using omnidirectional imaging. The technique involves training a CNN to perform room retrieval, addressed as an image classification problem. Additionally, the CNN is employed to embed the input image into a holistic descriptor from intermediate layers, aggregating relevant information that characterizes the input image. Additionally, we evaluate the influence of two main components on the localization performance: CNN architecture and effects applied in the data augmentation.

As for the CNN backbone, AlexNet shows excellent overall performance, especially when tested under the same lighting conditions than the training images. In contrast, ResNet performance decreases in sunny conditions which are the most challenging test conditions. This fact shows its low capability of generalization. The ResNext model surpass both in cloudy and sunny conditions, showcasing versatility across different lighting environments. However, EfficientNet exhibits a slight advantage over the ResNext model in terms of accuracy, although it requires more computational time. Furthermore, MobileNet consistently produces accurate results with a competitive computational time, demonstrating high performance across all conditions. Finally, the most striking result comes from ConvNext, which consistently achieves the highest accuracy in all scenarios, making it the top-performing model. Particularly noteworthy is its exceptional accuracy in sunny conditions, indicating its robustness and generalization capabilities.

Regarding the proposed data augmentation, training with the baseline dataset yields a remarkable accuracy, especially in cloudy and night conditions. However, a significant decrease is observed in sunny conditions, which diverge more from the training dataset. The spotlight effect shows marginal improvements, indicating that spotlight-based enhancement does not contribute to improve the generalization ability of the network. In contrast, data augmentation with shadows produces moderate improvements, especially in sunny conditions. Changing the overall brightness and darkness of the image produces substantial improvements, especially in sunny conditions. In addition, contrast-based effects are very effective, with significant improvements in all lighting conditions and especially in sunny conditions, improving results in this tough environment. Surprisingly, augmenting the dataset with changes in saturation shows a negative impact, especially in sunny conditions. Finally, increasing the dataset with rotations results in significant improvements in cloudy conditions. Finally, as for sunny conditions, the contrast effect yields the most optimal results, thereby enhancing the model's generalization capabilities and preventing overfitting.

In future works, studying more advanced techniques for generating more realistic visual effects with Generative Adversarial Networks (GANs) is a priority. Furthermore, we will evaluate other deep learning schemas such as Siamese, Triplet Neural Networks and Feature Pyramid Networks (FPNs). Finally, we will approach the localization problem in outdoor environments by using CNNs, considering the specificities of such scenarios.

Acknowledgements The Ministry of Science, Innovation and Universities (Spain) has supported this work through "Ayudas para la Formación de Profesorado Universitario" (FPU21/04969). This work is also part of the project TED2021-130901B-I00, funded by MCIN/AEI/10.13039501100011033 and the European Union "NextGenerationEU"/PRTR, and of the project PROMETEO/2021/075 funded by Generalitat Valenciana.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data availability Data is available in the github repo provided in Code availability.

Code availability Our code is publicly available on the project website https://github.com/juanjo-cabrera/IndoorLocalizationSingleCNN.git.

References

Aguilar WG, Luna MA, Moya JF, Abad V, Parra H, Ruiz H (2017) Pedestrian detection for UAVs using cascade classifiers with meanshift. In: 2017 IEEE 11th international conference on semantic computing (ICSC). IEEE, pp 509–514

Alfaro M, Cabrera JJ, Jiménez LM, Reinoso Payá L (2024) Hierarchical localization with panoramic views and triplet loss functions. arXiv preprint. arXiv:2404.14117

Bai D, Wang C, Zhang B, Yi X, Yang X (2018) CNN feature boosted SeqSLAM for real-time loop closure detection. Chin J Electron 27(3):488–499

Ballesta M, Payá L, Cebollada S, Reinoso O, Murcia F (2021) A cnn regression approach to mobile robot localization using omnidirectional images. Appl Sci 11(16):7521

Cabrera JJ, 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 Comput Sci 3(4):1–13

Cebollada S, Payá L, Jiang X, Reinoso O (2022) Development and use of a convolutional neural network for hierarchical appearance-based localization. Artif Intell Rev 55(4):2847–2874

Céspedes OJ, Cebollada S, Cabrera JJ, Reinoso O, Payá L (2023) Analysis of data augmentation techniques for mobile robots localization by means of convolutional neural networks. In: IFIP international conference on artificial intelligence applications and innovations. Springer, pp 503–514

Chen Z, Lam O, Jacobson A, Milford M (2014) Convolutional neural network-based place recognition. arXiv preprint. arXiv:1411.1509

Ding J, Chen B, Liu H, Huang M (2016) Convolutional neural network with data augmentation for SAR target recognition. IEEE Geosci Remote Sens Lett 13(3):364–368

Grisetti G, Stachniss C, Burgard W (2005) Improving grid-based slam with rao-blackwellized particle filters by adaptive proposals and selective resampling. In: Proceedings of the 2005 IEEE international conference on robotics and automation. IEEE, pp 2432–2437

Grisetti G, Stachniss C, Burgard W (2007) Improved techniques for grid mapping with rao-blackwellized particle filters. IEEE Trans Robot 23(1):34–46

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V (2019) Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1314–1324

Kim Y (2014) Convolutional neural networks for sentence classification. arXiv preprint. arXiv:1408.5882

Komorowski J, Wysoczańska M, Trzcinski T (2021) Minkloc++: lidar and monocular image fusion for place recognition. In: 2021 international joint conference on neural networks (IJCNN). IEEE, pp 1–8

Kopitkov D, Indelman V (2018) Bayesian Information Recovery from CNN for probabilistic inference. In: 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 7795–7802 . https://doi.org/10.1109/IROS.2018.8594506

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, p 25

LeCun Y, Bengio Y (1995) Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks. MIT Press, Cambridge

Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S (2022) A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11976–11986

Milford MJ, Wyeth GF (2012) Seqslam: visual route-based navigation for sunny summer days and stormy winter nights. In: 2012 IEEE international conference on robotics and automation. IEEE, pp 1643–1649

Naseer T, Ruhnke M, Stachniss C, Spinello L, Burgard W (2015) Robust visual SLAM across seasons. In: 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 2529–2535

Perez L, Wang J (2017) The effectiveness of data augmentation in image classification using deep learning. arXiv preprint. arXiv:1712.04621

Pronobis A, Caputo B (2009) COLD: COsy localization database. Int J Robot Res 28(5):588–594. https://doi.org/10.1177/0278364909103912

Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788

Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, p 28

Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp 234–241

Rostkowska M, Skrzypczynski 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

Salamon J, Bello JP (2017) Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process Lett 24(3):279–283. https://doi.org/10.1109/LSP.2017.2657381

Sarlin P, Cadena C, Siegwart R, Dymczyk M (2019) From coarse to fine: robust hierarchical localization at large scale. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 12708–12717 . https://doi.org/10.1109/CVPR.2019.01300

Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2013) Overfeat: integrated recognition, localization and detection using convolutional networks. arXiv preprint. arXiv:1312.6229

Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6(1):60

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint. arXiv:1409.1556

Sünderhauf N, Shirazi S, Dayoub F, Upcroft B, Milford M (2015) On the performance of convnet features for place recognition. In: 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 4297–4304

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9

Tan M, Le Q (2021) Efficientnetv2: smaller models and faster training. In: International conference on machine learning. PMLR, pp 10096–10106

Uy MA, Lee GH (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

Wang H, Yang W, Huang W, Lin Z, Tang Y (2018) Multi-feature fusion for deep reinforcement learning: sequential control of mobile robots. In: International conference on neural information processing. Springer, pp 303–315

Xie S, Girshick R, Dollár P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1492–1500

Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A (2014) Learning deep features for scene recognition using places database. In: Advances in neural information processing systems, pp 487–495

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.