







Analysis of Data Augmentation Techniques for Mobile Robots Localization by Means of Convolutional Neural Networks

Orlando José Céspedes¹, Sergio Cebollada^{1,2}(✉) , Juan José Cabrera¹ ,
Oscar Reinoso^{1,2} , and Luis Payá¹ 

¹ Institute for Engineering Research, Miguel Hernández University, Elche, Spain
orlando.cespedes@goumh.es,

{s.cebollada,juan.cabreram,o.reinoso,lpaya}@umh.es

² Valencian Graduate School and Research Network for Artificial Intelligence,
Valencia, Spain

Abstract. This work presents an evaluation regarding the use of data augmentation to carry out the rough localization step within a hierarchical localization framework. The method consists of two steps: first, the robot captures an image and it is introduced into a CNN in order to estimate the room where it was captured (rough localization). After that, a holistic descriptor is obtained from the network and it is compared with the descriptors stored in the model. The most similar image provides the position where the robot captured the image (fine localization). Regarding the rough localization, it is essential that the CNN achieves a high accuracy, since an error in this step would imply a considerable localization error. With this aim, several visual effects were separately analyzed in order to know their impact on the CNN when data augmentation is tackled. The results permit designing a data augmentation which is useful for training a CNN that solves the localization problem in real operation conditions, including changes in the lighting conditions.

Keywords: Mobile Robotics · Omnidirectional Vision · Hierarchical Localization · Deep Learning · Data Augmentation

1 Introduction

Artificial intelligence (AI) techniques have been commonly proposed to address computer vision and robotics problems. Among the existing techniques, Convolutional Neural Networks (CNNs) are one of the most popular to address a variety of problems. With the emergence of 360° vision sensors, the use of omnidirectional images has been widely proposed to address localization tasks in mobile robotics. Regarding the methods to extract relevant information from the images, the use of global-appearance descriptors has been extensively evaluated and the results show

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-34111-3_42.

this approach as a successful solution. Furthermore, recent works have proposed obtaining such holistic descriptors from intermediate layers of CNNs. For example, Aguilar et al. [1] propose a pedestrian detector for UAVs (Unmanned Aerial Vehicles) based on a combination of Haar-LBP features with Adaboost and cascade classifiers with Meanshift; Wang et al. [11], use an autoencoder for fusion and extraction of multiple visual features from different sensors with the aim of carrying out motion planning based on deep reinforcement learning.

Previous works [4] have proposed hierarchical visual models to carry out the localization task efficiently. This method consists in arranging the visual information hierarchically in different layers of information in such a way that the localization can be solved in two main steps. First, a coarse localization to roughly know in which room or area the robot is and second, a fine localization in this pre-selected area. CNNs have proved to be a successful technique in many practical applications. There are well known architectures which have been used as starting point to address new computer vision tasks. For instance, GoogLeNet was proposed by Szegedy et al. [10]. This network has 22 layers, it is trained for object classification but it uses 12 times fewer parameters than AlexNet. As for the use of CNN to solve localization tasks, Kopitkov and Indelman [5] propose using CNN holistic descriptors to estimate the robot position of learning a generative viewpoint-dependent model of CNN features with a spatially-varying Gaussian distribution. Sarlin et al. [8] carry out a hierarchical modelling using a CNN, which extracts local features and holistic descriptors for 6-DOF localization. The coarse localization is solved by using global retrieval and global descriptors and the fine localization is solved by matching local features.

Regarding the use of CNNs, a complete and varied training is essential, thus, a large training dataset must be available. Since a lack of a large enough dataset 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 apply visual effects that may occur in real operation conditions with the aim of making 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. For example, Perez and Wang [7] present a study about the effectiveness of the data augmentation to solve the classification task. Shorten and Khoshgoftaar [9] 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 which 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 in order to carry out a data augmentation for CNN training to address hierarchical localization. This work focuses on the rough localization step, which is solved by using the output layer of the CNN for room retrieval. Hence, the efficiency of each visual effect will be assessed through the ability of the CNN to robustly estimate the room where the image was captured.

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.

2 Methodology

2.1 CNN Adaption

Building a CNN from scratch to solve a specific task can be tough, since it requires of a deep expertise and also having a proper dataset to address the training. Moreover, as studied in previous works [2], adapting and re-training well-known networks for a different purpose can lead to accurate results. In this sense, the present work proposes departing from the Places CNN [12], which was trained for scene recognition. Places presents an architecture similar to AlexNet [6] and it was trained with around 2.5 million images to classify the candidate image among 205 categories of scenes. Figure 1 shows the architecture for a better comprehension.



Fig. 1. Architecture of the Places CNN [12]. This network was created to address a scene recognition task among 205 types.

This work proposes to train the CNN departing from panoramic images to address a room retrieval in an indoor environment composed of nine different rooms, hence, some layers of the original CNN are replaced. First, the input layer is re-adapted from $227 \times 227 \times 3$ to $128 \times 512 \times 3$. Second, the fully-connected layer fc8, softmax and output layer are replaced to fit them to the new classification task (scene recognition among 9 possible indoor rooms). After the replacement of those layers, the new CNN is trained to solve the rough localization (room retrieval).

2.2 Hierarchical Localization Approach

The aim of this work is to address visual localization by means of a hierarchical approach using deep learning as follows: first, a rough localization step is carried out to retrieve an area of the environment, and second, a fine localization step is tackled in that pre-selected area to refine the position fitting.

The output of the CNN is used to solve the rough localization step. In this sense, to train the CNN, a set of images that cover the target environment is captured, and each image has a label that indicates the room from which it was taken. With this information, the CNN is trained to solve the room retrieval problem. Once the CNN is correctly trained, the hierarchical localization is

solved as follows: a test image is introduced into the CNN and the output layer indicates the room where it was captured. Simultaneously, a holistic descriptor is obtained from an intermediate layer of the same CNN and this descriptor is compared with all the descriptors of images in the map. The nearest neighbour is retrieved and, then, the coordinates of the capture point of the retrieved image are considered the current position of the robot.

2.3 Data Augmentation Techniques

Training a model means establishing the parameters to address the desired task. Hence, if the model has a wide variety of parameters, the training process needs many examples. Often, the number of instances in the training process is small. In this sense, data augmentation is a good solution, because it is possible to avoid overfitting. This basically consists in creating new instances of ‘data’ by applying different visual effects. Apart from avoiding overfitting, considering visual effects that may occur in real operation conditions will make the CNN robust against those effects.

In previous works that train a model for visual localization [3] diverse effects were applied, such as orientation changes, reflections, general changes in illumination, noise, occlusions, etc., and it was proved that the use of this technique improves. These effects are applied over each image in the original dataset either individually or jointly. However, all the generated images are put together in a new augmented training dataset. Therefore, the influence of each kind of effect over the performance of the resulting CNN is not clear. The aim of this work is to apply different data augmentation effects individually, in order to assess their impact in the resulting CNN.

The present work focuses on two kinds of visual effects: changes of illumination conditions and changes of orientation. As for the changes of illumination conditions, we consider the following effects:

- **Spotlights and shadows:** Circular sources of light such as bulbs appear very often in indoor environments. Moreover, the presence of darker areas by object shades is also usual. Hence, it is proposed to increase the pixels values to simulate more light intensity (spotlight) and decreasing to simulate shadows (shadow spotlights). Position, shape of the spotlights and maximum values are randomly selected in order to consider different changes of illumination. Spotlights and shadow spotlights are applied separately for different data augmentation options.
- **General brightness and darkness:** The low intensity values of the original images are increased in order to create new images brighter than the original ones. This effect simulates a higher general level of illumination in the scene (for example, a sunny day). On the contrary, the high intensity values are decreased in order to create new images darker than the original ones. This effect simulates a lower light supply (for example, capturing the images at night). Brightness and darkness are applied separately, but used for the same data augmentation.

- **Contrast:** The contrast of the image plays an important role as it permits differentiating objects in the scene. Moreover, images with low contrast tend to have a smoother appearance with few shadows and reflections.
- **Saturation:** Color saturation refers to the color intensity given by the pixels. The less saturation, the less colorful is the image (even looking a gray-scale image for very low saturation). Such phenomenon may also occur in real environments and it is also considered in the data augmentation.

Concerning the orientation changes, they can happen during imaging capturing when the robot captures images in the same position but with a different orientation. Regarding this data augmentation option, for each original image, new ones are generated by applying a rotation of n , where $n = i \times 10^\circ$, $i \in [1, 35]$. Hence for each image, 35 additional images are generated.

3 Experiments

This section presents the results obtained through the use of the CNN to carry out the rough localization step. Concerning the training process, it consists in using either the original cloudy dataset (composed of 519 images) or the augmented (cloudy) dataset. Three main experiments are tackled. First, the use of a data augmentation based on orientation changes is evaluated. Second, each illumination effect is considered separately when training the CNN. Finally, a third experiment is developed whose data augmentation consists in applying jointly all the visual effects.

As for the rough localization step, once the model has been trained, the process is as follows: (1) the robot captures a new omnidirectional image from an unknown position within the environment. (2) The image is converted to panoramic. (3) The panoramic image is introduced into the CNN in order to estimate the most likely room where the image was obtained. The images used for training the model are not used to evaluate the localization task and three test datasets with different illumination conditions have been considered with the aim of evaluating the robustness of the CNN. Hence, the datasets implicated in the present work for testing the proposed approach are the following:

- Cloudy dataset, whose illumination is the same as the one used for training the CNNs. This dataset contains 2778 images (different from the images in the training dataset).
- Night dataset, whose images were obtained at night, hence, some areas present a considerable lack of light. This dataset contains 2707 images.
- Sunny dataset, whose images were captured during a sunny day, hence, the illumination is higher in general and the windows are also a source of light. This dataset contains 2807 images.

With the aim of validating the robustness of each resulting CNN, this work proposes the use of the accuracy metric. Moreover, it is also interesting to analyze the confusion matrix obtained for each test.

3.1 Experiment 1: Orientation Changes

As mentioned before, it is very likely that the orientation of the robot is not the same as the one presented during the mapping process, thus, the model should present robustness against changes of orientation. In this experiment, the data augmentation technique consisted in applying 35 different orientation changes over each of the the training images. After that, the CNN is trained and tested using the three test datasets. The results are shown in the Fig. 2 (a).

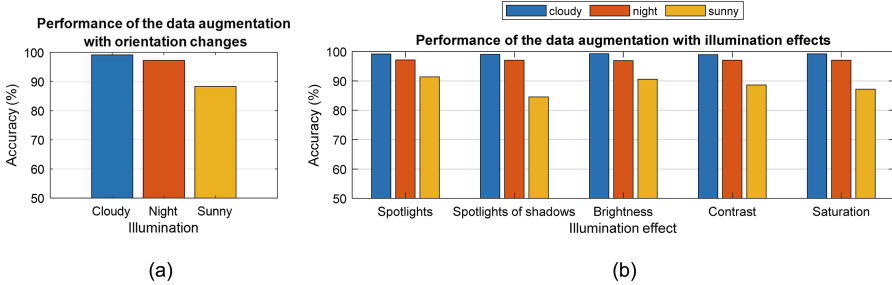


Fig. 2. Room retrieval results for DA with (a) orientation changes and (b) with illumination effects. Accuracy of the CNN to estimate the room where the test images were captured. Results presented under cloudy (blue), night (red) and sunny (yellow) illumination conditions. (Color figure online)

Regarding the performance under cloudy conditions (no change in the lighting conditions with respect to the training), the model reaches an accuracy of 99.17%. Nonetheless, the accuracy decreases when illumination changes are presented: 97.27% and 88.31% for night and sunny test datasets respectively. From these results, in case of using only orientation changes in the data augmentation, the conclusion is that the network retrieves the room without problems when there are no illumination changes, but it outputs worse results, specially at the presence of brighter images. Therefore, a profound study about illumination effects will be developed in the following experiment.

3.2 Experiment 2: Illumination Effects

This subsection tests the influence of illumination effects in detail in order to improve the performance of the CNN. Five effects are considered: spotlights, spotlights of shadows, general brightness/darkness, contrast and saturation. Every type of effect is applied individually over the training dataset to obtain each augmented dataset. Different effects are not blended with the aim of knowing the importance of each effect. For each image, five different levels of spotlights/contrast/saturation are applied. Regarding the brightness effect, for each image, three levels of brightness and three levels of darkness are applied.

The results for the different effects under the three illumination conditions are shown in the Fig. 2(b).

Concerning the results under the night illumination conditions, the best accuracy is obtained with the ‘spotlights’ effect (97.16%) and the worst accuracy with the ‘brightness/darkness’ effect (96.9%). However, the difference between those results is not significant. As for the performance under the sunny illumination conditions, the accuracy is lower than with the night dataset. However, it is remarkable that the accuracy has improved in comparison with the results obtained with the ‘orientation changes’. The best accuracy is obtained with the spotlights effect (91.38%) and the worst case (Saturation, 87.17%) is similar to the accuracy obtained with the ‘orientation change’ data augmentation.

If the accuracy results for the test datasets are considered in average, (see Table 1) we can see the weight of each visual effect concerning the rough localization.

Table 1. Classification of the illumination effects regarding their average accuracy (considering jointly the cloudy, night and sunny test datasets)

Visual effect	Average accuracy (%)
1. Spotlights	95.90
2. Brightness	95.57
3. Contrast	94.86
4. Saturation	94.48
5. Spotlights shadows	93.54

Finally, Fig. 3 shows the accuracy obtained by using a CNN which was trained without data augmentation and the computing time required to train all the evaluated models. Since the models were trained with different numbers of images, this variable is normalized by dividing the computing time by the total number of images used to train each network.

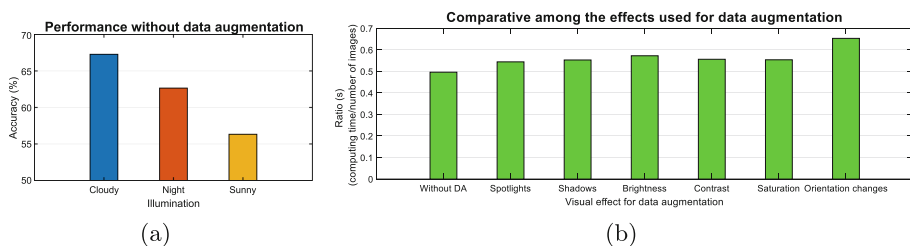


Fig. 3. (a) Accuracy of the CNN to estimate the room without applying data augmentation and (b) normalized computing time for training the neural network.

In general, all the evaluated data augmentation techniques lead to a higher accuracy than using the CNN trained only with raw data. Concerning illumination changes, the six proposed methods improve the accuracy. It should be pointed out that the data augmentation technique based in orientation changes provides relatively good results under night and sunny illumination conditions, reaching the best accuracy when evaluating the night dataset. As for the accuracy under sunny illumination conditions (the most challenging ones), ‘spotlights’ (91.38%) and ‘brightness’ (90.59%) are the effects that provide more than 90% of accuracy. Regarding the computing time, there are not significant differences between the methods, since the fastest one (without data augmentation) requires 0.5 s per image and the slowest (with orientation changes) needs 0.65 s per image. Moreover, this process is offline, thus computing time is not as crucial as accuracy. As for the computing time with illumination effects, all proposed DA techniques present similar values (around 0.55 s per image).

3.3 Experiment 3: Evaluation of the Data Augmentation Considering All the Effects Jointly

The last evaluation concerning this study consisted in carrying out the training of the CNN by using a data augmentation using jointly all the effects. The aim of this experiment is to analyze whether the mix of effects can lead to a better room retrieval performance. The training of this neural network was similar to the previous ones, but the number of epochs was reduced to 20, because by that epoch, the training and validation accuracy stopped increasing and more epochs would have led to overfitting. The data augmentation consisted in considering a unique augmented dataset that contains the images created for the six previous data augmentation techniques and no images with more than one visual effect were created for this purpose. The results under the three illumination conditions are shown in the Fig. 4. This figure shows a comparison between three CNN models: without data augmentation, with data augmentation based in ‘spotlights’ effects and with data augmentation based in all the studied effects (spotlights, spotlights of shadow, brightness/darkness, contrast, saturation and orientation changes).

This figure shows that the results for the data augmentation based in all effects present better accuracy than without data augmentation and similar than using only ‘spotlights’ data augmentation. As for sunny illumination conditions, the accuracy gets worse than using ‘spotlights’ data augmentation.

Apart from the accuracy, the study of the confusion matrices can lead to more insightful conclusions. Figure 5 shows an example of the confusion matrix obtained under night illumination conditions. These results were obtained by doing room retrieval using CNN with data augmentation applying all effects. This figure shows that the room confusion is only given between next-door rooms. For example, corridor presents more incorrect predictions because it is connected to the most of the places. Figures of confusion matrix without data augmentation and with data augmentation applying all effects under the three illumination conditions can be found in the Electronic Supplementary Material. From these

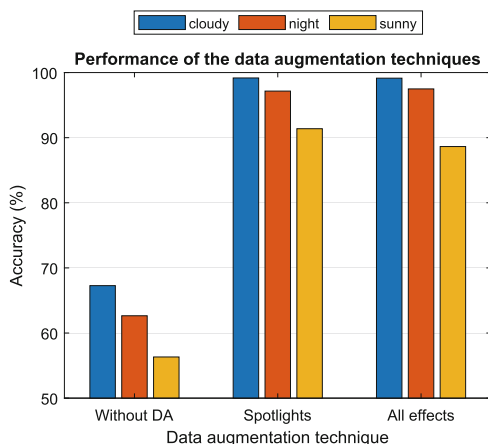


Fig. 4. Accuracy of the CNN for room retrieval under cloudy, night and sunny illumination conditions. The models were trained (left bars) without data augmentation, (center bars) with data augmentation with 'spotlights' effects and (right bars) with all studied effects.

figures, considering sunny illumination as the worst scenario and using data augmentation, the produced errors are relatively low and can be controlled by reinforcing the algorithm as it was proposed in previous works. For instance, using likelihood thresholds to select more than one room as candidate when the first retrieved room is not confident enough [4].

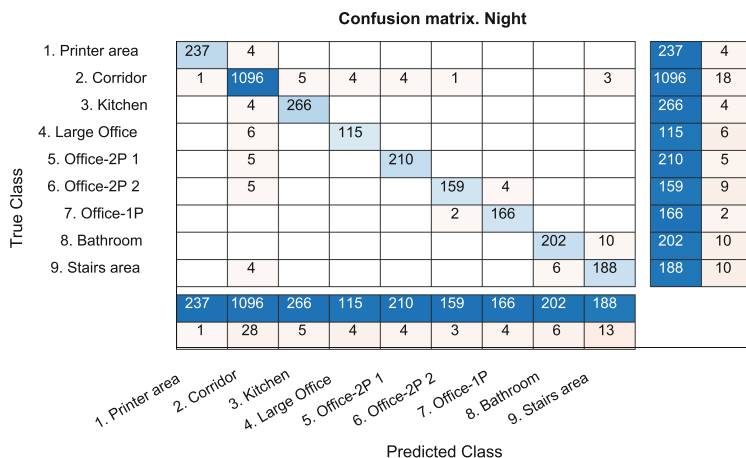


Fig. 5. Confusion matrix under night illumination conditions. CNN trained with data augmentation.

4 Conclusions

In this work, we have evaluated a set of visual effects to carry out a data augmentation for CNN training. The objective of the CNN is to address a room retrieval task, which is used as rough step within a hierarchical localization approach. The evaluated visual effects were considered because they may appear in images when the localization task is addressed. Hence, these effects should be considered by the network during the training process in order to be capable of coping with them. This study has focused on two kinds of visual effects: orientation and illumination changes, which frequently happen when the environment is revisited by the robot. With the aim of analyzing the influence of each effect, they were evaluated separately. After that, all effects are considered jointly in a unique training dataset.

In general, all the considered DA techniques provide benefits for the room retrieval task. The accuracy is substantially higher by using a CNN trained with any data augmentation technique than without it. As for the data augmentation with orientation changes, it shows that the accuracy is improved in general in comparison without performing data augmentation. It should be noted that this technique also improved the results concerning changes of lighting conditions, which may not be expected at the beginning of this study.

Regarding the illumination effects, ‘spotlights’ was able to successfully simulate the visual effects that are present both under night conditions, whose main light sources are in the upper part of the environment, and sunny conditions, whose lighting sources are located in the upper part of the environment and in the windows/picture windows. ‘Spotlights of shades’ was the least suitable effect to train the neural network for room retrieval purposes. This may be due to a lower importance of shadows from the point of view of global appearance or because the simulation of this effect does not suit the presence of shadows as properly as spotlights do with lighting sources.

Despite the worse results are given under sunny illumination conditions, it is notable that the best improvements are obtained under this condition if we compare the accuracy using data augmentation and without it. In this sense, it should be also considered that in indoor environments, images captured at night can be more similar to those captured during a cloudy day, because the rooms keep a similar level of light intensity (a level suitable to address work activities). The main illumination in the night dataset is obtained from lighting sources placed on the ceiling, as in the cloudy dataset. However, in the sunny dataset, the main lighting sources are different when there are picture windows. Moreover, lighting reflections are presented on the floor. These visual differences explain the reason why the sunny illumination condition is more challenging regarding global appearance.

When all effects are used to address the data augmentation, the results improve at night and get worse for sunny conditions. This is pointing out that effects, such as the darkness produced by ‘brightness/darkness’ effect and the orientation changes’ effect, allow to improve the accuracy. Concerning the worsening under sunny conditions in comparison with the ‘spotlights’ data augmentation,

this can indicate that the ‘spotlights’ effect has lost its importance within the training dataset. Hence, to avoid it, a weighting regarding the number of images per effect could be explored.

In future works, we will focus on developing a method to quantify the weight for each visual effect for data augmentation. We will also extend our study to other effects such as occlusions produced by furniture changes and the presence of objects in front of the camera. Also, we will study more in detail how to include illumination effects that suit better the night and sunny conditions and we will study other deep learning approaches such as transformers or networks using attention mechanism.

Acknowledgements. This work has been supported by the ValgrAI (Valencian Graduate School and Research Network for Artificial Intelligence) and Generalitat Valenciana. This work is also part of the project PID2020-116418RB-I00 funded by MCIN/AEI/10.13039/501100011033, and of the project PROMETEO/2021/075 funded by Generalitat Valenciana.

References

1. Aguilar, W.G., Luna, M.A., Moya, J.F., Abad, V., Parra, H., Ruiz, H.: Pedestrian detection for UAVs using cascade classifiers with meanshift. In: 2017 IEEE 11th International Conference on Semantic Computing (ICSC), pp. 509–514. IEEE (2017)
2. Ballesta, M., Payá, L., Cebollada, S., Reinoso, O., Murcia, F.: A cnn regression approach to mobile robot localization using omnidirectional images. *Appl. Sci.* **11**(16), 7521 (2021)
3. Cabrera, J.J., Cebollada, S., Flores, M., Reinoso, Ó., Payá, L.: Training, optimization and validation of a cnn for room retrieval and description of omnidirectional images. *SN Comput. Sci.* **3**(4), 1–13 (2022)
4. Cebollada, S., Payá, L., Jiang, X., Reinoso, O.: Development and use of a convolutional neural network for hierarchical appearance-based localization. *Artif. Intell. Rev.* **55**(4), 2847–2874 (2022)
5. Kopitkov, D., Indelman, V.: Bayesian information recovery from CNN for probabilistic inference. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 7795–7802 (2018). <https://doi.org/10.1109/IROS.2018.8594506>
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012)
7. Perez, L., Wang, J.: The effectiveness of data augmentation in image classification using deep learning. arXiv preprint [arXiv:1712.04621](https://arxiv.org/abs/1712.04621) (2017)
8. Sarlin, P., Cadena, C., Siegwart, R., Dymczyk, M.: 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 (2019). <https://doi.org/10.1109/CVPR.2019.01300>
9. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. *J. Big Data* **6**(1), 60 (2019)
10. Szegedy, C., et al.: Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9 (2015)

11. Wang, H., Yang, W., Huang, W., Lin, Z., Tang, Y.: Multi-feature fusion for deep reinforcement learning: sequential control of mobile robots. In: Cheng, L., Leung, A.C.S., Ozawa, S. (eds.) ICONIP 2018. LNCS, vol. 11307, pp. 303–315. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04239-4_27
12. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Advances in Neural Information Processing Systems, pp. 487–495 (2014)