ENS 491-492
Graduation Project
Final Report

Project Title:
Indoor Localization Using Camera Images

Group-127 Members:

Ayça Elif Aktaş

Mustafa Mert Gökbayrak

Supervisor(s):

Mustafa Ünel

Date: 26.05.2024

1. EXECUTIVE SUMMARY

Our project addresses the challenge of indoor localization for mobile robots, where traditional methods like GPS and static landmarks are often ineffective due to signal interference and environmental changes. We developed an advanced indoor localization system using computer vision and deep learning techniques, specifically employing convolutional neural networks (CNNs) such as PoseNet and the AlexNet architecture trained on Microsoft's 7-Scenes dataset. This system accurately classifies rooms, estimates poses, and calculates coordinates, enabling dynamic navigation that adapts to environmental changes. Our methodology involved several key stages: selecting and utilising large datasets to train deep learning models, refining these models, and integrating them into a coherent system for localization. We employed a two-stage pipeline approach for localization, where the AlexNet model was used for room classification, followed by PoseNet models for pose estimation, tailored for each specific scene. This approach ensured high precision in determining each frame's location. The results demonstrated that our deep learning-based solution performs comparably to conventional localization techniques while offering enhanced accuracy in indoor environments. The AlexNet model achieved high accuracy in room classification. The PoseNet models, tested with both resnet18 and resnet50 architectures, showed varied performance across different scenes, with resnet50 generally performing better in terms of lower translation and rotation errors.When compared to the benchmark established by Kendall et al. (2015), our models showed competitive performance, with improvements in translation error for most scenes. The final pipeline combining AlexNet and PoseNet models achieved an average translation error of 0.4215 metres and an average rotation error of 11.5806 degrees, indicating a significant advancement over existing technology.

2. PROBLEM STATEMENT

A major gap in the rapidly developing field of automation technology is highlighted by the difficulty of interior localization for mobile robots, particularly in complex and dynamic indoor situations. In these environments, traditional localization techniques like GPS and static landmark-based systems often fail. When indoor infrastructure blocks GPS signals, the resulting localization data is either unavailable or erroneous. Because they depend on predetermined, static settings, static landmarks are unable to accommodate dynamic changes like moved furniture or changed layouts that are typical

in sectors like logistics and healthcare. Moreover, these approaches suffer from inadequate lighting or a lack of distinguishing visual markers in the environment, which makes navigation more difficult for autonomous robots. This results in lower operating effectiveness, more navigational errors, and higher accident risks.

Inspired by the significant growth in the worldwide robotics industry, particularly in the area of industrial automation, our research seeks to improve upon the shortcomings of existing localization systems through the application of deep learning and computer vision technologies. Our system uses cameras to dynamically navigate indoor spaces by incorporating advanced image processing algorithms. Specifically, we used Microsoft's 7-Scenes dataset to develop convolutional neural networks (CNNs) for pose estimation and image categorization. By using visual data to continuously update and improve the robot's comprehension of its environment, this methodology far outperforms previous approaches and allows for more accurate navigation and task performance. Strict respect for technical and scientific standards during system design ensures outstanding performance, dependability, and safety. It takes into account important elements that are necessary for reliable operation in a variety of interior environments, including frame rate, pose, and the translation of the robot and camera resolution. Our method intends to revolutionise mobile robot navigation and localization by utilising state-of-the-art deep-learning techniques in computer vision, thereby establishing new standards for scalability and flexibility in robotic navigation technology. In addition to addressing a crucial aspect of mobile robot navigation—localization in interior environments—this project advances the field by offering a novel, practically sound, and highly flexible response to an enduring problem.

2.1 Objectives/Tasks

The objectives and their intended results from the start of the project to the end are listed as follows in chronological order:

i. Research on Indoor Localization:
   Finding the gaps in the present approaches and possible deep learning innovation areas.

ii. Research on Computer Vision Basics:
   Thorough comprehension of the techniques used in computer vision for robot localization and navigation.

iii. Research on Utilising Deep Learning on the Topic:

Selecting the best deep learning models to increase the precision and effectiveness of localization.

iv. Research on AlexNet Algorithm:

Evaluation of AlexNet's performance using the Microsoft 7-Scenes dataset for indoor scene recognition and classification.

v. Research on IEEE Standards Related with the Project:

Commitment to all important IEEE standards to guarantee system security and dependability.

vi. Image Processing and Feature Extraction Method Learning: Efficient indoor image feature extraction that enables additional processing and analysis.

vii. Depth and Distance Calculations using the Extracted Features: Precise distance and item location estimation, improving the robot's awareness of its environment.

viii. Project Initialization and Setup:

A development environment with good planning that facilitates effective experimentation and iteration.

ix. Model Training and Initial Testing with AlexNet:

Initial evaluation of the model's ability to identify and classify indoor environments.

x. Experimental Design Testing:

Optimised experimental setup that maximises the efficiency and success of the processes used to train and test the models.

xi. Comparison of Different PoseNet Models and Optimizers:

Choosing the best model version and optimizer configuration to improve the accuracy of pose estimation.

xii. Optimization of the Model:

A highly optimised model that performs indoor localization tasks with exceptional reliability and accuracy.

xiii. Comparative Analysis and Project Wrap-up:
Comprehensive documentation and analysis that discusses the project's accomplishments, advances, discipline, and identifies potential study topics for future.

2.2 Realistic Constraints

Economic Constraints:
The selection of hardware components and computer resources are essential for implementing deep learning techniques—were greatly impacted by financial constraints. Throughout the implementation process, we gave priority to the most economical solutions; we used open-source software and reasonably priced, dependable hardware to avoid the exorbitant expenditures of proprietary technology. The project's budget was adhered to by spending since the development process was meticulously planned to maximise resource allocation.

Computational Resource Constraints:
Training time and model complexity are strongly impacted by the computational power available, particularly the GPU and CPU capabilities, which have a significant impact on the performance of our project's deep learning models. We balanced the computational demand and available resources by optimising neural network topologies and training protocols in order to handle this.

Engineering and Scientific Standards:
To guarantee the system's performance, dependability, and integration capability, industry standards have to be followed. We made sure that our design parameters complied with international standards for performance, safety, and interoperability by adhering to IEEE camera standards (*IEEE standard on Video techniques: Measurement of resolution of camera systems, 1993 techniques*) and other appropriate guidelines for imaging systems.

3. METHODOLOGY

Our project is strategically designed as a two-stage pipeline to effectively localise a frame within indoor environments. This pipeline approach ensures that each stage of localization is handled efficiently and accurately, reflecting a sequential flow of data processing that enhances overall system performance. The initial stage of our pipeline involves the use of the AlexNet model (Krizhevsky et al., 2017) for room classification.

Here, all test data frames are first processed through the AlexNet model. This model plays a critical role as the primary classifier that determines the specific room or scene from which each frame originates. Following the room classification, the pipeline progresses to the second stage involving pose estimation. Based on the room identified by the AlexNet model, an appropriate PoseNet model is dynamically selected and invoked. Each PoseNet is specialised for its respective scene, having been trained exclusively with data from that particular environment. This specialisation enables the PoseNet models to perform highly accurate pose estimation by leveraging their tailored understanding of the specific spatial and visual features of the room. This structured pipeline approach underlines the project's innovative strategy in tackling the complexities of indoor localization by decomposing the problem into manageable, sequentially dependent tasks. Through this methodology, our system ensures that each frame's location is determined with high precision.

Room Classification and Pose Estimation

The methodology for our research combines a variety of advanced computational techniques and tools to handle the issues of indoor localization. This section describes the step-by-step approach used, from model and dataset selection to specific methods during the training and testing phases. The basis of our computational strategy is based on Convolutional Neural Networks (CNNs), notably the AlexNet architecture (Krizhevsky et al., 2017) for room classification and several PoseNet (Kendall et al., 2015) models for position estimation. AlexNet was chosen because of its strong track record in image classification tasks, which is crucial for the first phase of determining the room where a frame is positioned. The architecture's capacity to cope with complicated image data and extract significant features makes it appropriate for our application. For posture estimation, we used PoseNet18 and PoseNet50 models that were adjusted to each scene in the dataset. At first the PoseNet18 model is used to train for all the 7 scenes in the dataset, the training and testing is completed and recorded for the PoseNet18 models. After PoseNet18, to use a much bigger model for our research PoseNet50 is chosen to better optimise the results. This decision was motivated by the need for high precision when establishing the frame's position and orientation in a room. Each PoseNet model was trained independently for each scene to guarantee that it could effectively learn and anticipate the unique spatial properties of each environment.

Dataset

As the dataset to train and test the PoseNet and AlexNet models in the project Microsoft's RGB-D 7-Scenes (*RGB-D dataset 7-scenes,* 2013) dataset is used, which is a comprehensive collection of RGB and depth photos from diverse indoor scenes, complete with position matrices for each frame. This dataset was useful in training and testing our models since it contains a wide range of indoor circumstances, including varying angles and lighting scenarios, which are critical for determining the robustness of our localization algorithms. The dataset consists of randomly taken pictures of 7 different rooms named office, heads, pumpkin, stairs, fire, red kitchen, and chess. All of these scenes are shown with respect to the same order, with a sample image from the dataset in the below Figure 1. The dataset's structure is stated in the Microsoft's related website (*RGB-D dataset 7-scenes,* 2013) as follows: a screenshot of the raycasted dense reconstruction, TrainSplit.txt and TestSplit.txt files to determine the sequences to be used in the training and testing parts of the models. Each scene in the dataset consists of multiple sequences that each includes 1000 frames in it. These sequences also include these for each of the 1000 frames separately: the pose files that are in the form of a 4x4 matrix, a depth PNG showing the depth in millimetres, and lastly a colour PNG, which shows the frame in RGB.

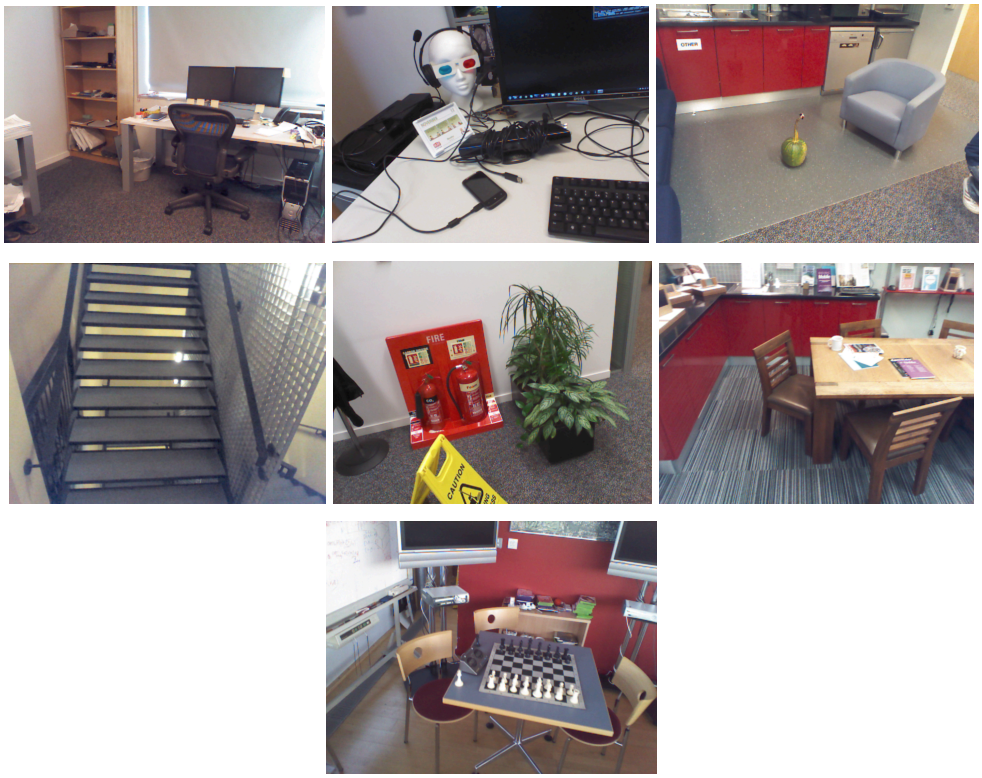


Figure 1: Sample images from the Microsoft's RGB-D 7-Scenes Dataset

Training

Both the AlexNet and PoseNet models were initialised with pre-trained weights, which used transfer learning to accelerate the training process and increase model performance without requiring large computational resources. This method allowed us to begin with a solid foundation, as these models had previously learnt robust feature representations from large-scale image datasets. AlexNet was trained to classify rooms using images from all seven scenes. The goal of training the AlexNet model on images from all seven scenes is to enable it the capacity to generalise across various interior contexts. By exposing AlexNet to a diverse range of circumstances, we improve its ability to correctly identify any given input frame to one of the predefined scenes. This stage is critical since the conclusion directly affects the upcoming posture estimation method. Each PoseNet model was trained individually using data from the associated scenario. This scene-specific training was critical for achieving high posture estimate accuracy because

each scene has its own set of spatial characteristics. While training We used a 5-fold cross-validation approach for both the AlexNet and each PoseNet model. This method divides the dataset into five unique subsets, guaranteeing that each fold acts as the test set once and the remaining four folds as the training set. Training and validation for each model were repeated across all five folds to reduce any bias caused by random data splitting and improve model generalisation. Each model was trained for 15 epochs per fold, which gave the networks enough time to learn from the data without overfitting. This constant training approach across all folds guarantees that each model is extensively evaluated under diverse settings, offering a thorough evaluation. We used the CrossEntropyLoss criterion for the AlexNet model, which is responsible for room classification, because it is effective at dealing with multi-class classification challenges. The model was optimised using Stochastic Gradient Descent (SGD), with a learning rate of 0.001 and momentum of 0.9. This combination is very successful at efficiently converging to an optimal solution, which improves the model's capacity to generalise across varied indoor scenes. Similarly, all seven PoseNet models tasked with pose estimation employed the Mean Squared Error (MSELoss) as a criterion, and the Stochastic Gradient Descent optimizer, with the same settings as used for the AlexNet model. This option is appropriate for regression assignments since it calculates the average of the squares of the errors, or the average squared difference between the estimated and actual values. The same optimizer, SGD, with settings of 0.001 learning rate and 0.9 momentum, was utilised. This assures uniformity in how each model approaches minimising loss, allowing for a fair comparison of performance across scenes. In the Figure 2 below the pipeline that will be applied for all test images in the test dataset for the project is represented, and the method for the room classification and pose estimation is also indicated.

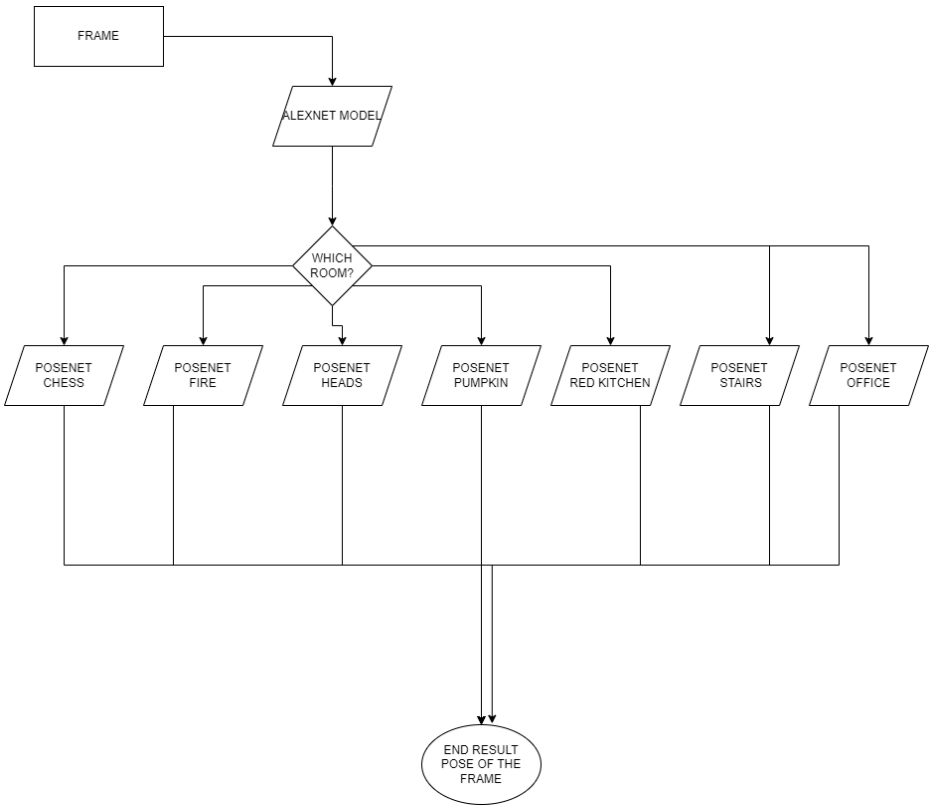


Figure 2: The project pipeline for room classification and pose estimation

The training and validating of the AlexNet model is completed with the obtained validation loss of 0.01% from the Fold 2 out of 5 folds. The validation and training losses are shown in the plot in the Appendix section. Following training, the models underwent rigorous testing to assess their performance. The AlexNet room classification model was tested using a combined set of test data from all scenes, demonstrating its ability to assign every given frame to the appropriate scene. In contrast, each PoseNet model was tested separately using test data particular to each scenario, allowing us to examine the accuracy of posture prediction in a controlled setting. The performance of our models was evaluated using several metrics. For room classification, accuracy was the primary metric, indicating the percentage of test images correctly classified to their corresponding scene. For pose estimation, we used more granular metrics such as the average error in estimated pose compared to the ground truth, which includes metrics like mean squared error (MSE) in translation and rotation. After validation and training, the model iteration

that displayed the highest accuracy and lowest error rates on the test data after the cross-validation process was chosen for the PoseNet model and AlexNet model. This technique assures that we choose the most effective model configuration for practical applications, hence increasing the indoor localization system's reliability and efficiency.

4. RESULTS & DISCUSSION

Fold	Accuracy	Validation Loss
0	98.94%	0.0001
1	99.00%	0.0001
2	99.11%	0.0001
3	98.97%	0.0001
4	99.10%	0.0001
Averaged Model	98.13%	N/A

Table 1: AlexNet test dataset accuracy results across 5 folds.

Scene	Validation Loss	Average Translation Error	Average Rotation Error
Chess	0.42%	0.3235	9.18
Fire	0.36%	0.3680	17.54
Heads	0.40%	0.2996	13.46
Office	0.39%	0.3404	8.88
Pumpkin	0.47%	0.7224	19.35
Red Kitchen	0.44%	0.5102	10.31
Stairs	0.49%	0.4261	11.87

Table 2: Average Translation Error is in metres and Average Rotation Error is in degrees. Performance of the best PoseNet models (resnet18) on the test data for various scenes.

Scene	Validation Loss	Average Translation Error	Average Rotation Error
Chess	0.39%	0.2632	7.4473
Fire	0.30%	0.3868	18.5901
Heads	0.33%	0.2727	14.2431
Office	0.37%	0.3402	8.8737
Pumpkin	0.45%	0.7062	19.5365
Red Kitchen	0.39%	0.4521	10.2388
Stairs	0.41%	0.4518	12.2715

Table 3: Average Translation Error is in metres and Average Rotation Error is in degrees. Performance of the best PoseNet models (resnet50) on the test data for various scenes.

Scene	Average Translation Error (m)	Average Rotation Error (degrees)
Chess	0.32	8.12
Fire	0.47	14.4
Heads	0.29	12.0
Pumpkin	0.47	8.42
Red Kitchen	0.59	8.64
Stairs	0.47	13.8
Office	0.49	7.68

Table 4: The benchmark established by Alex Kendall, Matthew Grimes, and Roberto Cipolla in their 2015 paper, ”PoseNet: A Convolutional Network for Real-Time 6- DOF Camera Relocalization.”

	Average Translation Error of all 7 scenes (metres)	Average Rotation Error of all 7 scenes (degrees)
Our Work	0.4215	11.5806
The Benchmark	0.4428	10.4371

Table 5: Performance of the final pipeline combining the AlexNet model and PoseNet models for all 7 scenes on test data compared to the averaged out result of the benchmark across all 7 scenes.

The AlexNet model performed consistently throughout all five validation folds, with accuracy values ranging from 98.94% to 99.11% (Table 1). The averaged model accuracy was significantly lower (98.13%), but it still demonstrates strong performance. The validation loss remained consistent at 0.0001 across all folds, demonstrating the model's robustness and stability in room categorization tasks. This degree of accuracy shows that the AlexNet model is quite successful at discriminating between different rooms in an indoor environment.

The PoseNet models, tested with both resnet18 and resnet50 architectures, performed differently in different indoor settings (Tables 2 and 3). The average translation and rotation errors are important metrics for determining the model's accuracy in calculating the camera's position and orientation. Both the ResNet18 and ResNet50 models produced low validation loss values across different scenes, with the resnet50 performing somewhat better overall. The ResNet18 model had an average translation error of 0.2996m to 0.7224m, whereas the ResNet50 model varied from 0.2632m to 0.7062m. Notably, the ResNet50 model produced less translation problems in all of the scenes except the "Fire" and "Stairs" scenes than the resnet18 model. The rotation errors for the resnet18 model varied from 8.88° to 19.35°, whereas the ResNet50 model ranged from 7.44° to 19.53°. The ResNet50 model performed better on estimating the rotation in all of the scenes except the "Fire", ” Pumpkin” and "Stairs" scenes which had worse rotation errors on the ResNet50 model than the ResNet18 model.

The provided 3D camera trajectory plots (Figures B1-B7 in the Appendix) visualise the predicted versus ground truth camera trajectories for different scenes ("fire," "pumpkin," "heads," "red kitchen," "stairs," "chess," and "office") using the PoseNet50 model.

"Stairs" Scene (Figure B1)

The "stairs" scene shows significant deviations between the predicted and ground truth trajectories. The blue points are widely dispersed, indicating substantial errors in the model's pose estimations. This suggests that the PoseNet50 model struggles with the vertical and horizontal variations in this scene, leading to higher average translation errors.

"Chess" Scene (Figure B2)

For the "chess" scene, the predicted trajectory closely follows the ground truth, with minor deviations. The model performs well in this environment, indicating that the PoseNet50 model can accurately estimate poses where the spatial layout is relatively simpler and more consistent.

"Fire" Scene (Figure B3)

In the "fire" scene, the predicted trajectory (blue) diverges significantly from the ground truth (red) in several areas. The PoseNet50 model shows a higher concentration of predictions in the middle of the trajectory even though the ground truth trajectory is not mainly concentrated in the middle area, indicating potential issues with pose estimation in this specific environment. The larger spread of the blue points that do not follow the red path suggests that the model struggles with maintaining accuracy over this scene's translation predictions.

"Heads" Scene (Figure B4)

For the "heads" scene, the model's predicted trajectory shows a moderate deviation from the ground truth. The prediction points are scattered more widely, particularly towards the start and end of the trajectory. This indicates that while the PoseNet50 model can generally follow the correct path, it has difficulties in accurately predicting positions at the trajectory boundaries.

"Office" Scene (Figure B5)

In the "office" scene, the predicted trajectory again shows a reasonable alignment with the ground truth. However, there are noticeable deviations, especially in areas with dense point concentrations. The model exhibits better performance in the central regions of the scene in the trajectory but struggles at the boundaries, leading to errors in pose estimation.

"Pumpkin" Scene (Figure B6)

The "pumpkin" scene's trajectory indicates that the model predictions do not closely follow the ground truth, the predicted path that is noted with blue has some notable deviations. The predicted points are densely clustered around the centre of the scene, suggesting that the PoseNet50 model predicates correctly the frames that are around the central region of the scene but struggles at predicting the frames at the boundaries, leading to errors in pose estimation.

"Red Kitchen" Scene (Figure B7)

In the "red kitchen" scene, the predicted trajectory aligns fairly well with the ground truth. This scene seems to be handled better by the PoseNet50 model compared to others, but there are still areas where pose estimation errors are evident.

The trajectory plots illustrate that the PoseNet50 model's translation performance varies

significantly across different scenes. While the model shows promise in scenes like "chess", “redkitchen”, and “office” it struggles in more complex environments like "stairs”,”pumpkin and "fire." The results from these visualisations also align with the results that we found for “average translation error” for all PoseNet models with resnet50 (Table3). These visualisations highlight the importance of scene-specific optimizations and suggest areas for further improvement in the pose estimation process. Reducing these deviations would likely involve enhancing model robustness to handle the diverse spatial characteristics present in different indoor environments.

Our models' translation and rotation errors are competitive when compared to the Kendall et al. (2015) benchmark (Table 4). Our models performed significantly better in terms of translation error in all of the scenes except the "Pumpkin" scene where our translation errors are higher in both resnet18 and resnet50 models when compared to the Kendall et al. (2015) benchmark (Table 4). However, in case of rotation error all of the scenes except "Chess" and "Stairs," had worse results meaning higher error rates in our models when compared to the Kendall et al. (2015) benchmark (Table 4). Overall, for “Chess“ and “Stairs” scenes we produced better results in both translation and rotation error but for all other scenes we were able to only produce better result in only one field out of two criteria. In all of the scenes there is still room for improvement, especially in terms of rotation error.

The project has been completed with the implementation of the pipeline to obtain the final results on our test dataset. This implementation allowed us to validate our research hypotheses and quantify the overall performance of our models under various conditions. The final pipeline, which combined the best-performing AlexNet model with the best-performing PoseNet models, produced an average translation error of 0.4215 metres and an average rotation error of 11.5806 degrees on our test dataset for all seven scenes (Table 5). Our project has achieved a better “Average Translation Error rate for all 7 scenes” (Table 5), indicating an improvement over the existing technology when compared to the benchmark. The improvement over the Kendall et al. (2015) benchmark suggests that our method represents a viable innovation in indoor localization technology. These findings confirm the effectiveness of our combined technique for providing dependable and precise indoor localization.

5. IMPACT

Our work has had notable effects on science and technology while making substantial contributions to the field of indoor localization. The project has made scientific progress in applying deep learning techniques in complicated situations. Specifically, it has done so by applying AlexNet architecture-based convolutional neural networks and PoseNet for pose estimation to indoor localization problems. This has expanded our knowledge of how computer vision and robotic navigation may work together to increase the capabilities of mobile robots in indoor environments.

A significant technological advancement has been made with the creation of a reliable indoor localization system that outperforms conventional GPS and landmark based techniques. For mobile robots, this system's real-time adaptation to dynamic changes in indoor settings significantly improves navigation accuracy and operational efficiency. These developments are critical for use in industries like healthcare and warehousing, where accurate and dependable interior navigation is essential.

While the economic or entrepreneurial parts of our study were not explicitly investigated, the technologies produced could find future use in a variety of businesses that depend on advanced indoor navigation systems. The project's results may have an impact on how logistics operations are optimised and how well service robots perform in challenging conditions.

Regarding Freedom-to-Use (FTU) concerns, our project has complied with the use of open-source software and patent-free approaches, guaranteeing that the results are free from disputes arising from intellectual property. This strategy promotes an atmosphere of open innovation by enabling the project's outcomes to be freely used for future scholarly study and possible technical implementations without encountering legal limitations.

6.ETHICAL ISSUES

Our deep learning and computer vision indoor localization project involves some ethical questions, mostly related to data protection and privacy. The system is designed to process only visually relevant data for navigation inside preset contexts, hence reducing the possibility of unintentional personal data gathering via continuous imaging. All things considered, we work hard to make sure that our solution functions ethically, keeping a close eye on it and making changes as needed to handle any new ethical concerns.

7. PROJECT MANAGEMENT

Our study began with a focus on robot localization, with the goal of using a dataset particularly acquired from the perspective of a robot to address navigation and recognition issues in dynamic environments. However, as the project developed and we dug deeper into our study, we discovered some constraints and opportunities that required a shift in our strategy. Significantly, we switched from a robot-specific dataset to Microsoft's 7-Scenes dataset, which was recorded with a portable Kinect camera. This transition signified a shift from focusing primarily on robot localization to embracing a broader use of general camera frame localization. This decision was motivated by the need for more detailed data, the larger relevance of the findings, and the need to align our work with widely accepted benchmarks in the scientific community.

8. CONCLUSION AND FUTURE WORK

Our project achieved significant strides in enhancing indoor localization accuracy, particularly in reducing the Average Translation Error across most scenes when compared to benchmarks established in the study by Kendall et al., 2015. This improvement indicates the effectiveness of our methodology and the potential of our approach to refine the application of machine learning models in real-world scenarios. Despite these achievements, our study faces certain limitations. The primary constraint lies in the generalizability of our models, as they were specifically tailored for the 7-Scenes dataset and may not perform as effectively in environments that differ significantly from those included in this dataset. Additionally, while we managed to reduce translation errors, our rotation error metrics still require improvement to enhance overall localization precision. The results underscore the potential of specialised deep

learning models in improving spatial awareness in camera-based systems. However, further work is needed to address the limitations related to model generalizability and rotation accuracy. Looking ahead, there are several promising directions for advancing our project. Transitioning from basic transfer learning methods to fine-tuning methods may increase model accuracy by refining the models more deeply for our specific dataset and its unique properties. Also Using the Adam optimizer instead of Stochastic Gradient Descent (SGD) for pose estimation could provide insight into optimization strategies that may give superior results. Comparing the effectiveness of these two optimizers would aid in determining the most effective strategy for our models, as well as potential broader applications in indoor localization.

9. APPENDIX

Appendix A: Validation and Training Losses of the AlexNet Model from the 5-fold Training And Validation

The following figures illustrate the validation and training losses for our AlexNet model in 5 different folds. The orange lines in the graph represent the validation loss, and the blue line represents the training loss.

Figure A1. Training and Validation Losses of AlexNet Model for Fold 0:

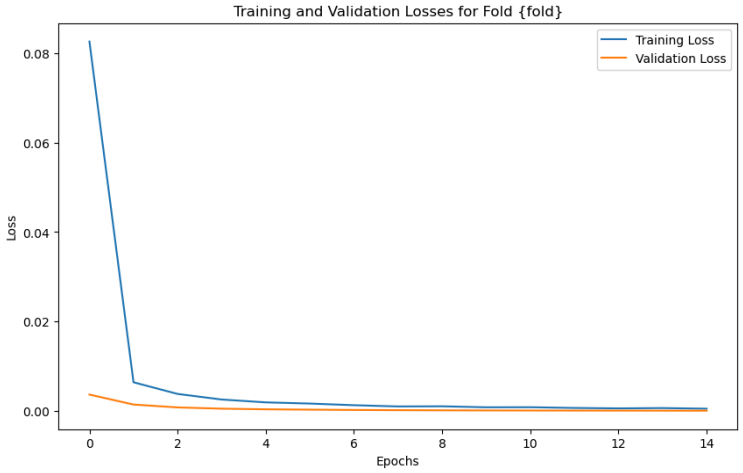


Figure A2. Training and Validation Losses of AlexNet Model for Fold 1:

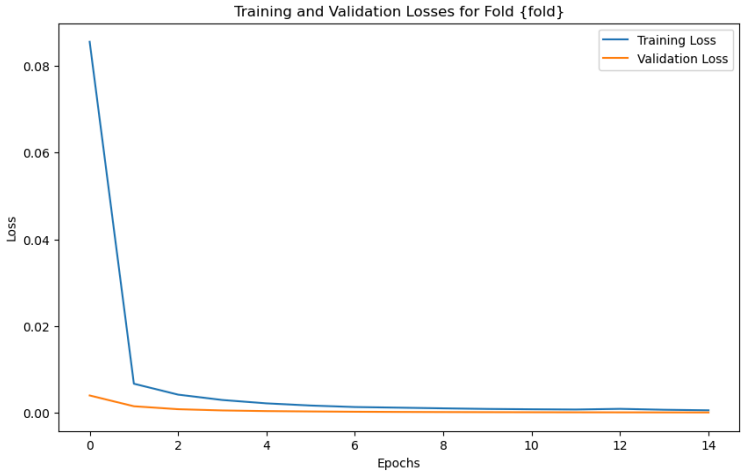


Figure A3. Training and Validation Losses of AlexNet Model for Fold 2:

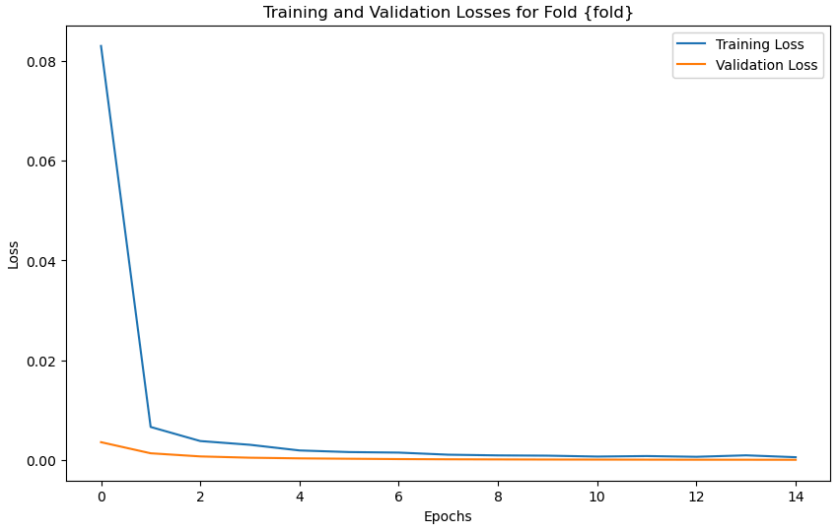


Figure A4. Training and Validation Losses of AlexNet Model for Fold 3:

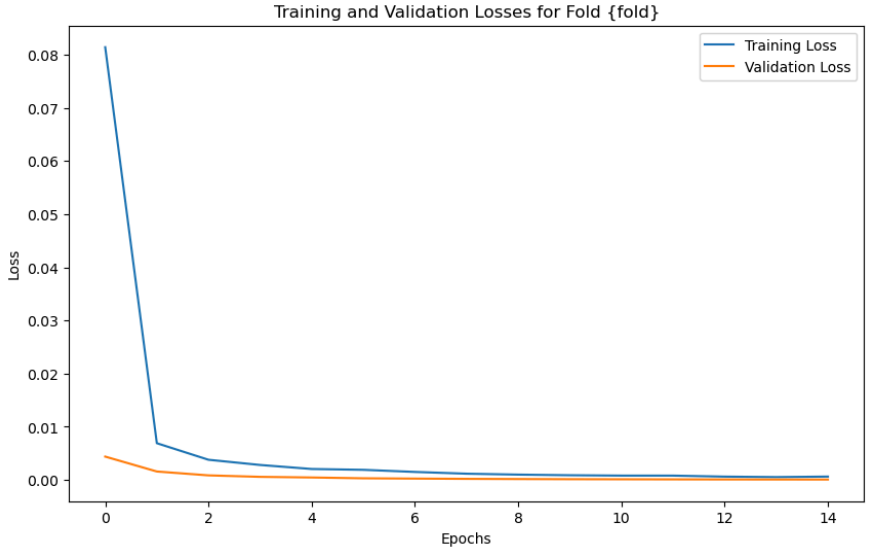
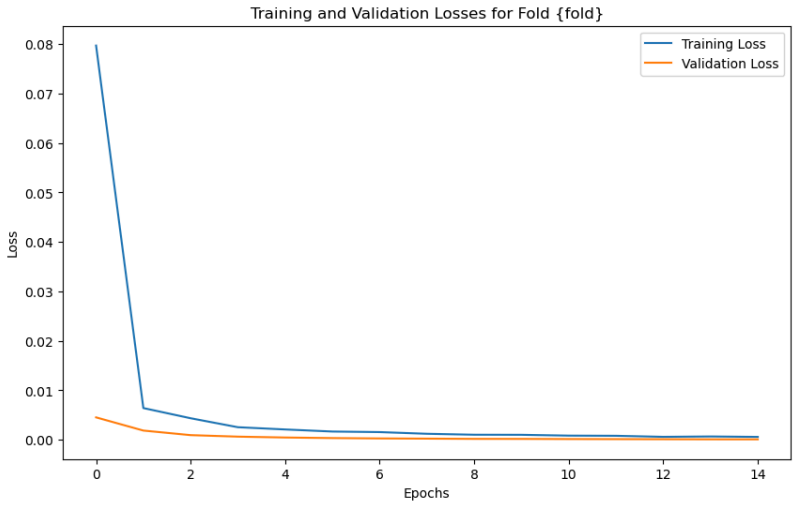


Figure A5. Training and Validation Losses of AlexNet Model for Fold 3:



Appendix B: 3D Camera Trajectories

The following figures illustrate the 3D camera trajectories for various scenes using the PoseNet50 model. The red lines represent the ground truth coordinates of each frame in the test dataset, depicting the actual trajectory followed by the camera. The blue points represent the predicted coordinates (translation) of each frame as determined by our PoseNet50 model on the test dataset.

Figure B1. 3D Camera Trajectory for "stairs" scene:

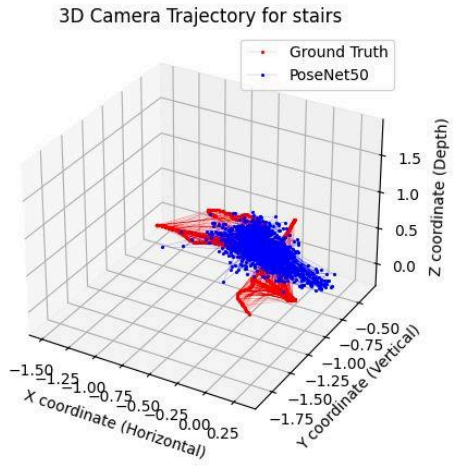


Figure B2. 3D Camera Trajectory for "chess" scene:

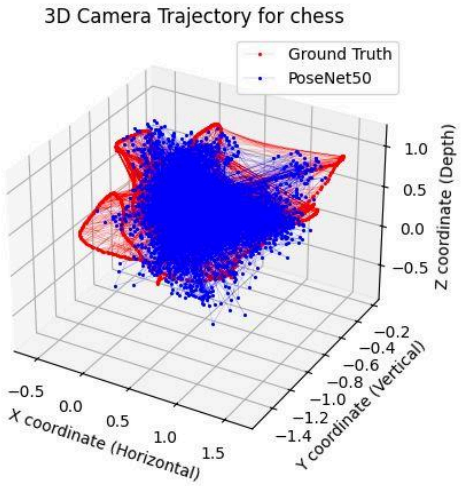


Figure B3. 3D Camera Trajectory for "fire" scene:

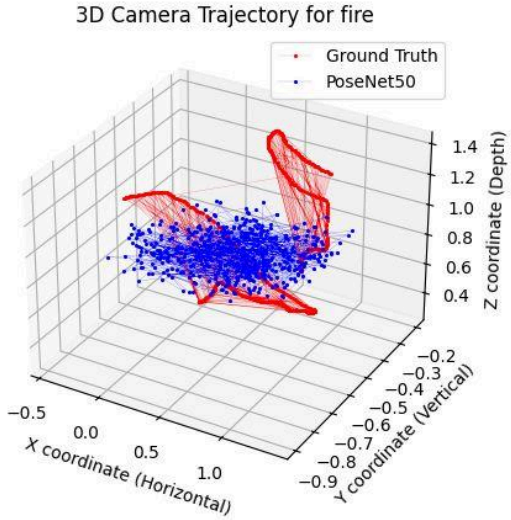


Figure B4. 3D Camera Trajectory for "heads" scene:

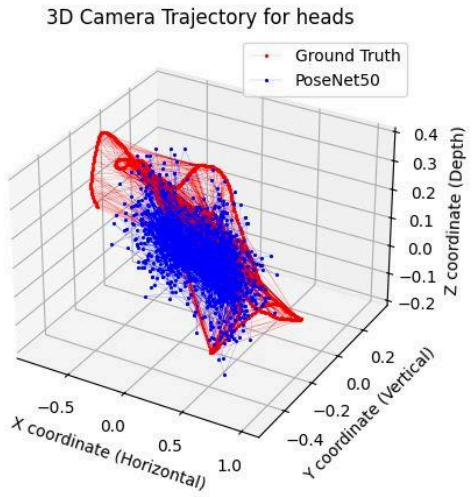


Figure B5. 3D Camera Trajectory for "office" scene:

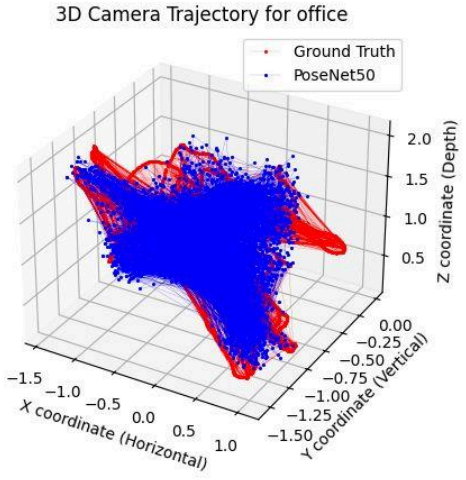


Figure B6. 3D Camera Trajectory for "pumpkin" scene:

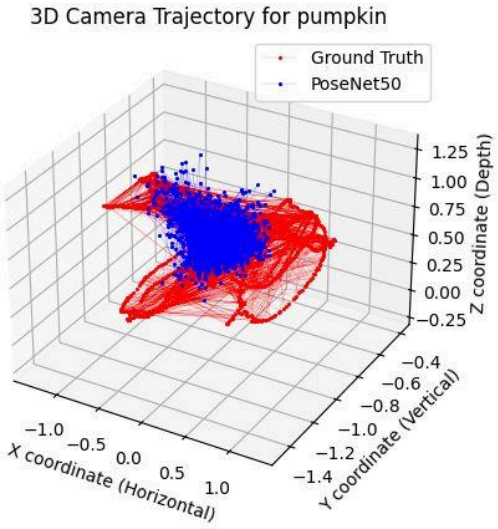
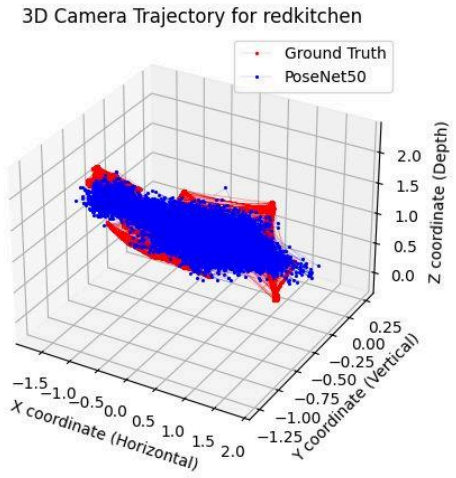


Figure B7. 3D Camera Trajectory for "redkitchen" scene:



10. REFERENCES

IEEE Standard on Video Techniques: Measurement of Resolution of Camera Systems, 1993 Techniques. https://doi.org/10.1109/ieeestd.1995.122626

Kendall, A., Grimes, M., & Cipolla, R. (2015). PoseNet: A convolutional network for real-time 6-DOF camera relocalization. *2015 IEEE International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv.2015.336

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

RGB-D dataset 7-scenes. Microsoft Research. (2022, September 7). https://www.microsoft.com/en-us/research/project/rgb-d-dataset-7-scenes/

Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., & Fitzgibbon, A. (2013). Scene coordinate regression forests for camera relocalization in RGB-D images. *2013 IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2013.377