ENS 492 – Graduation Project (Implementation)

Progress Report II

Project Title: Indoor Localization Using Camera Images

Group Number: 128

Group Members: Kağan Kağanoğlu, Yarkın Alpmen Akyosun

Supervisor(s): Mustafa Ünel

Date: 31.03.2024

1. PROJECT SUMMARY

1.1 Description:

This project focuses on developing an innovative solution for precise indoor localization within large enclosed spaces, such as shopping malls and airports, by leveraging computer vision and deep learning techniques. It aims to propose and evaluate different designs to compare with state-of-the-art approaches on different criteria, such as accuracy and speed.

Traditional methods like GPS and Wi-Fi face limitations indoors due to signal obstruction, prompting the need for a more effective solution, particularly those leveraging computer vision-based approaches.

1.2 Gap in the Literature:

Despite the significant advancements in indoor localization with computer vision technologies, there exists a notable gap that our project aims to address. While various studies have contributed to the field, including those by Fusco and Coughlan on indoor localization using computer vision and visual-inertial odometry [1], Niu and Li's work on automated image-based localization (HAIL) [2], and Akal et al.'s approach for single-platform image-based localization using distributed sensing [3], a comprehensive solution that effectively integrates computer vision-based approaches with deep learning techniques is still lacking.

While discriminative feature-based localization methods have been explored [4], and two-stage architectures involving image retrieval and pose estimation have shown promise [5], there remains a need for innovative methodologies that enhance both accuracy and efficiency.

While some studies have utilized deep belief networks [6] and convolutional neural networks [7] for image-based indoor localization, there is room for further refinement and advancement in these techniques to achieve higher levels of performance.

Additionally, we aim to explore and incorporate more modern approaches that may not be directly in the field of indoor localization but have proved useful in related fields, such as Swin Transformers [8] and You Only Look Once (YOLO) [9] object detection, to further push the boundaries of indoor localization accuracy and efficiency.

1.3 Motivation:

The motivation behind this project stems from recognizing the challenges faced by existing indoor location systems, particularly in large spaces such as shopping malls and airports, where conventional methods like GPS and Wi-Fi often fall short due to signal limitations. There is a clear need for more effective solutions.

The aim is to improve indoor localization technology, making it more reliable and efficient for users. By addressing the shortcomings of current approaches and exploring new methodologies, the intent is to enhance accuracy and efficiency.

1.4 Objectives and Intended Result:

Primary objectives of the project include, selecting suitable algorithms, and designing an advanced indoor localization system that meets realistic constraints such as accuracy, hardware requirements, and response time after conducting a thorough literature review. It is aimed to implement and test multiple designs, evaluate their performance, and select the most effective solution for practical application. The intended results encompass the design and implementation of a robust indoor localization system capable of accurate positioning in various indoor environments.

1.5 Fundamental Elements:

The project adheres to several fundamental elements essential for the successful development and implementation of an indoor localization with computer vision model. These elements ensure that our approach is thorough, systematic, and capable of producing reliable results.

Objectives: The project begins with clearly defined objectives that guide our efforts throughout the design process. These objectives encompass throughout literature search, identification of suitable algorithms, the selection of relevant datasets, and the development of innovative methodologies to address the challenges in indoor localization.

Criteria: Criteria is established to evaluate the performance and effectiveness of the proposed solution. These criteria encompass accuracy, efficiency, robustness, and scalability, ensuring that our model meets the requirements for practical application in various real-world scenarios.

Synthesis: The synthesis phase involves the integration of diverse methodologies, algorithms, and techniques to construct a cohesive indoor localization system. This process entails combining computer vision-based approaches, deep learning models, and modern technologies to enhance the capabilities of our design.

Analysis: Rigorous analysis is conducted to assess the performance and effectiveness of our proposed solution. This involves evaluating the accuracy of localization results, analyzing computational efficiency, and identifying potential areas for improvement or optimization while comparing the results with existing solutions.

Construction: The construction phase involves the implementation and development of the indoor localization system based on the synthesized design. This includes coding algorithms, integrating relevant libraries, and configuring hardware devices necessary for system operation.

Testing: Thorough testing is conducted to validate the functionality and performance of the constructed system. This includes conducting experiments using real-world benchmarking datasets, evaluating system outputs against ground truth data, and assessing the system's performance under various conditions and environments.

Evaluation: Comprehensive evaluation is performed to assess the overall effectiveness and reliability of the developed indoor localization system. This involves comparing system outputs with established benchmarks, and identifying areas for future research and improvement.

By adhering to these fundamental elements, the project aims to develop an advanced indoor localization model that is accurate, efficient, and adaptable to diverse indoor environments.

2. SCIENTIFIC/TECHNICAL DEVELOPMENTS

2.1 Object Detection Testing with YOLO

Motivation:

For design B, our initial focus was on implementing object detection using YOLO (You Only Look Once) [9], a foundational model capable of identifying objects within the indoor environment. YOLO detects objects present in the scene and provides essential information, including the object category and bounding box coordinates.

Method and Testing:

To enhance the accuracy of object boundary predictions, we integrated zero-shot learning into our approach. Zero-shot learning enables our system to generalize to new classes of objects, even in the absence of direct training data. This approach proved efficient for our case, as there was no need for training time, and YOLO demonstrated its capability to create bounding boxes even for unseen data.

During the testing phase, YOLO demonstrated efficiency in capturing bounding boxes around objects present in the scene. While the model excelled at providing bounding box predictions, its accuracy in correctly identifying objects was limited. However, since our primary objective was to create accurate bounding boxes, the incorrect identification of objects is irrelevant to our case.

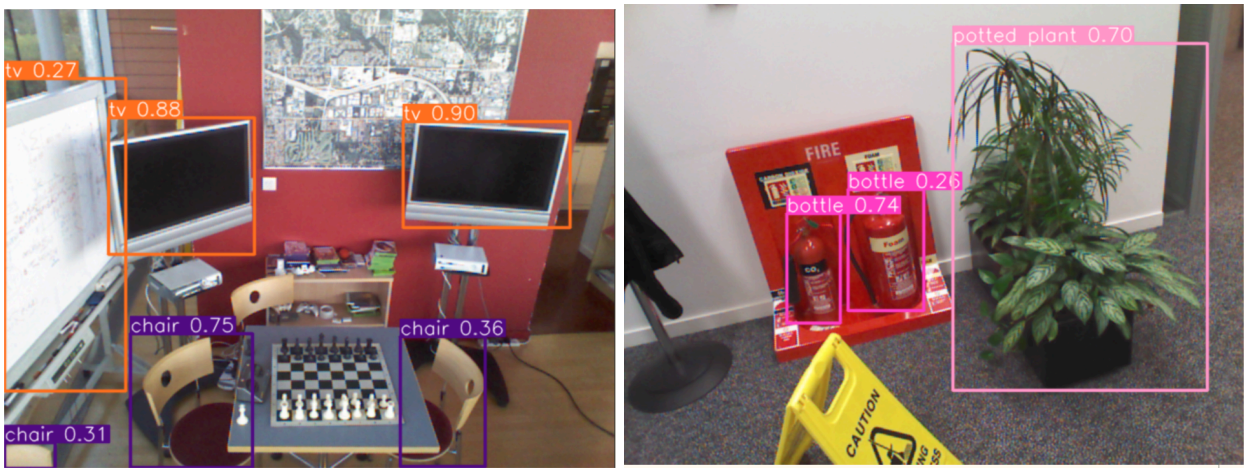

Figure [1]: YOLO fails to classify some objects, but generates overall good bounding boxes for most objects.

Future Steps:

Moving forward, the relative position estimation of object projections in 2D colored photos (bounding boxes) obtained from YOLO will serve as crucial input for the development of the model we plan to train. This model aims to predict position and orientation vectors, representing a critical step toward achieving accurate indoor localization.

2.2 Depth Estimation with Midas

Motivation:

As shown in our previous reports, both of our designs rely on depth estimation, which will be obtained from a model that can infer depth information from colored image. After our literature view, we have identified one such model called Midas. Midas is a monocular depth estimation model created by Microsoft. We planned on placing the model in our pipeline, however we decided to first put it under some preliminary tests to validate its robustness.

Definition and Hypothesis:

A “depth estimator” in our context is a model which can predict the distance of each pixel on an image from a camera device, by only relying on colored pixel values. Depth estimator will return an image of the same width and height where values represent distance. Therefore by definition, the ratio between depth values assigned to any two points in the input image must be equal to the ratio of their true distance from the camera. In other words, for

every image, there exists a linear scaling function which will map all predictions to real distance values correctly.

Naturally any depth estimation method in reality would not be perfect. But the above definition provides an empirical basis for evaluation of depth estimator's robustness. Since training is a time and resource consuming process, we wanted to validate midas architecture's robustness by using a pretrained model provided by creators of midas to assess its zero-shot performance on our dataset.

Test Setup and Results:

We evaluated the performance of pretrained midas models by calculating error rates between the ground truth depths and the predicted images, which were scaled using a fitted linear regression. As stated earlier, an ideal depth estimator would yield zero error in this scenario for all images. However, for non-ideal estimators, estimation errors cause a difference between ground truth and fitted scale.

7-Scenes dataset contains sequences of paths sampled with the Kinect device. Each sequence is composed of subsequent frames and each frame is a tuple of a rgb image, a 16-bit depth image and homogenous coordinates of the device. Depth images contain the distance of each pixel from the device in units of millimeters.

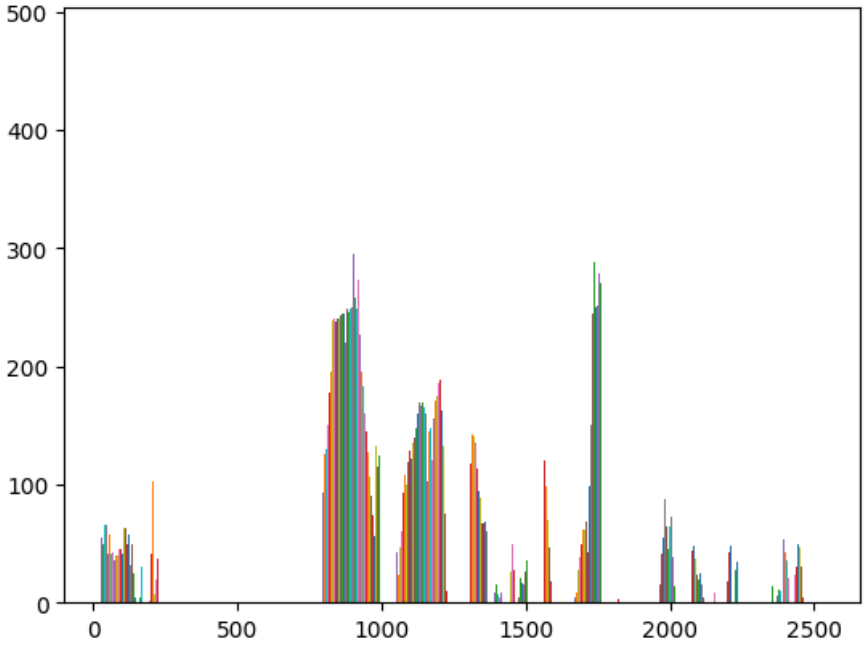

Figure [2]: Histogram of a depth image. (in millimeters)

For every sample image, scale fitting is done by applying linear regression to regress grand truth depths from predicted depth values. An important detail to note is that depth images are 16 bits while midas predictions are 8 bit, therefore midas predictions are first converted to 16 bits before regression. After the scale is determined, it is applied to the original predicted image to obtain a scaled prediction. Absolute difference between scaled prediction and grand truth provides valuable insight for midas models assessment. Figure [3] demonstrates absolute error between ground truth and scaled prediction depths (unit is millimeters). Error histograms show that most of the error is concentrated in lower values. Blue and red lines show mean and median error respectively. An important observation regarding error histograms is that spaced “spikes” of error are presented, which indicate different regions of image with varying degrees of estimation error. Therefore depth estimation is likely to be more accurate for certain objects and less accurate for others. But we are currently unable to test this hypothesis since we don’t have a robust object detection model yet.

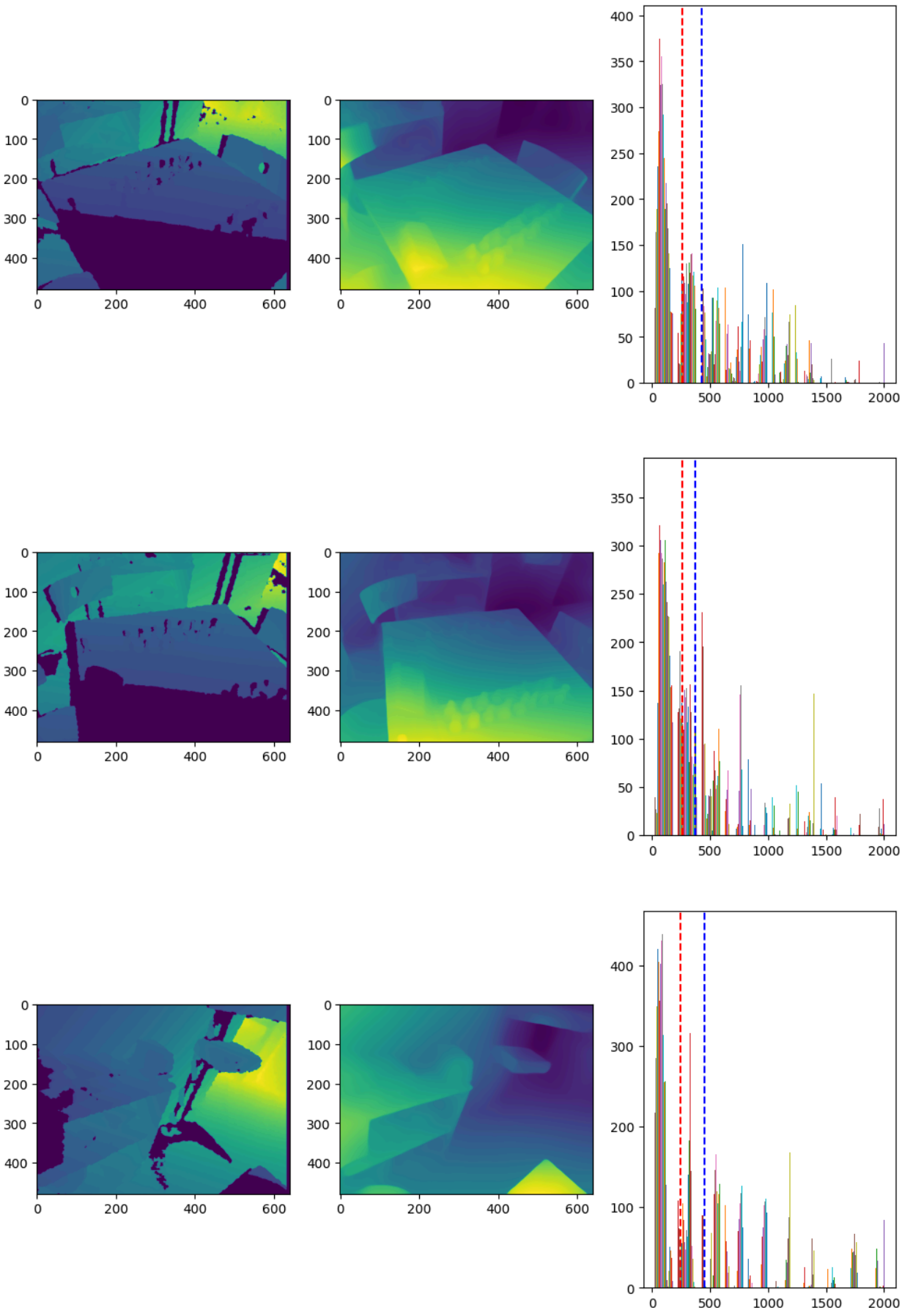

Figure [3]: From left to right, depth image, scaled prediction, histogram of absolute difference between them

We performed tests for 4 of the training sequences in our dataset. 50 images are randomly pulled from each sequence. To ensure selection homogeneity and reproducibility, selection is made with uniform random function with seed values. For all tested images mean and median error values are collected and placed in below histograms. We observe that error rates are relatively concentrated, with around %3 outliers with very high levels of error (which are not visible in histograms range). Another interesting observation is that error rates for “fire” room is slightly lower than that of the chess room.

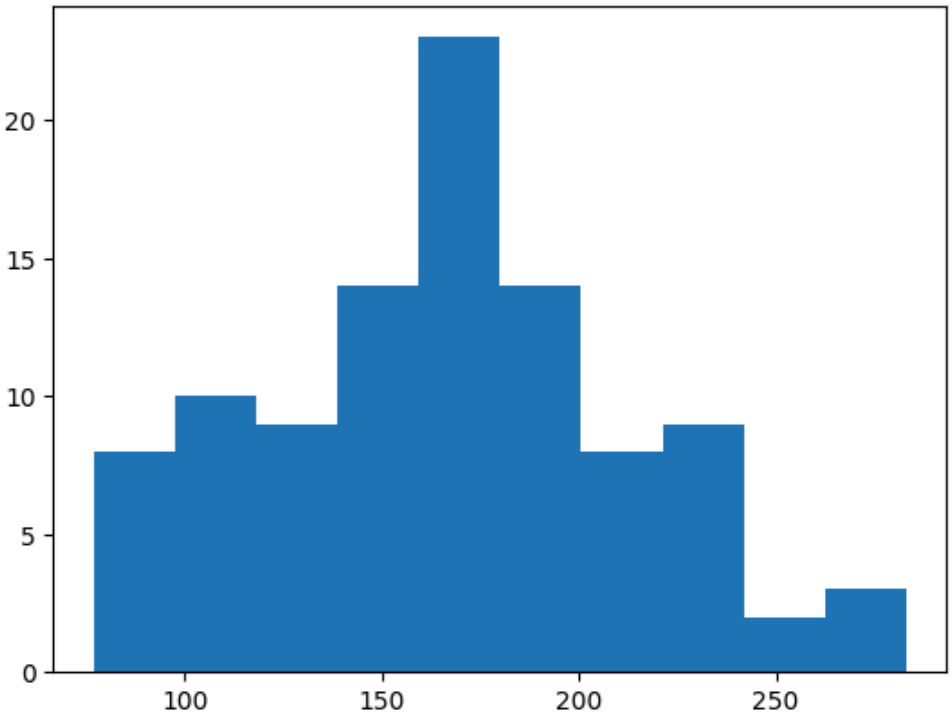

Figure [4] : Fire sequence 1 & 2, median error histogram (in millimeters)

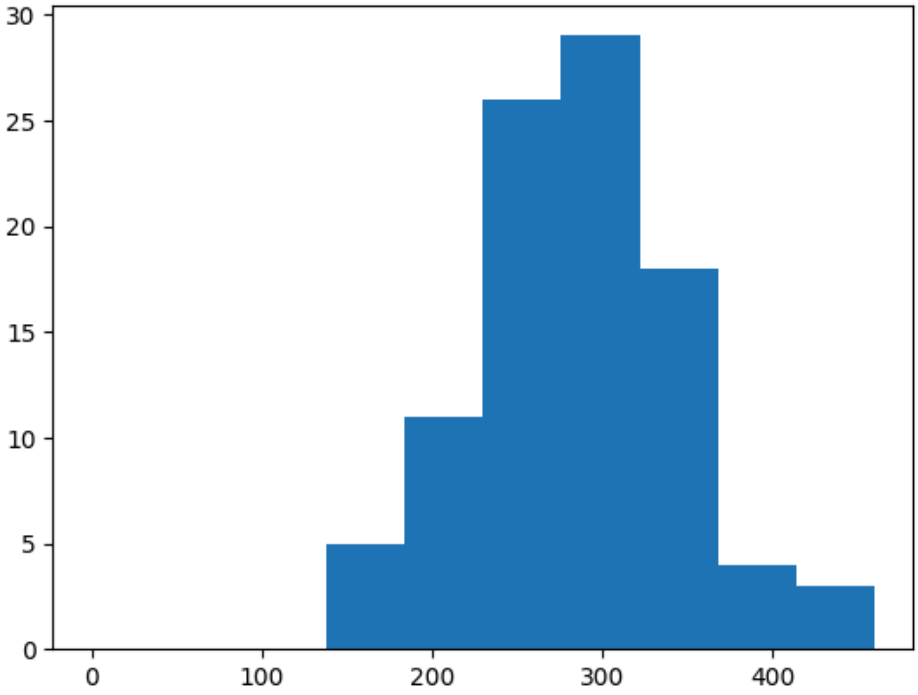

Figure [5] : Fire sequence 1 & 2, mean error histogram (in millimeters)

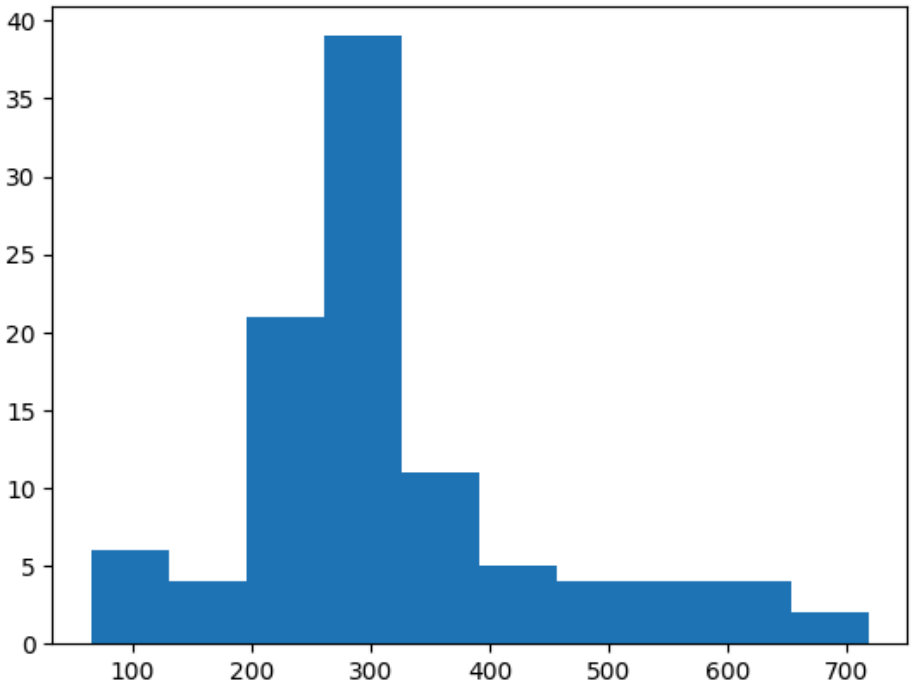

Figure [6]: Chess sequence 1 & 2, median error histogram (in millimeters)

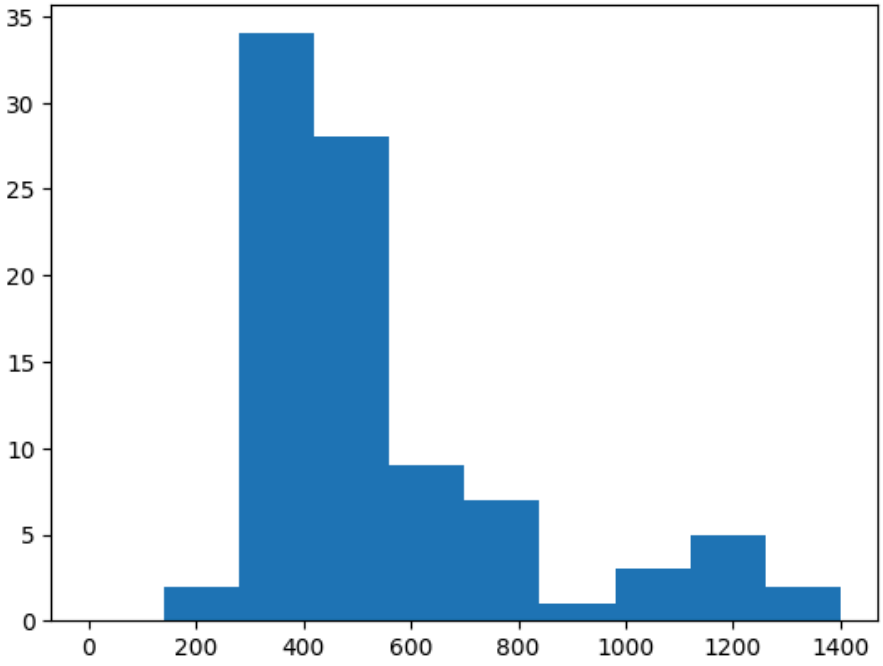

Figure [7] : Chess sequence 1 & 2, mean error histogram (in millimeters)

Experiments we have conducted yield relatively low absolute error levels for the majority of the test samples, indicating robust depth estimation performance. Validation of depth estimation performance was important for subsequent steps of our project since forward steps of our architecture depend on depth estimation to accurately infer position.

3. ENCOUNTERED PROBLEMS

Changes in Project Plans and Goals:

During the progression of the project, a change was made in the design approach for Design A. Originally, a self-attention layer was planned for incorporation into the neural network architecture. However, upon further research and consideration, we decided to replace this layer with Swin Transformer from the paper Swin Transformer: Hierarchical Vision Transformer using Shifted Windows by Ze Liu et al. [8]. This decision was driven by the recognition of Swin Transformer's superiority in terms of modernity, power, and potential for enhancing the

accuracy and efficiency of indoor localization. Swin Transformer has shown promising results on tasks of recognizing indoor images, which is believed to be relevant to our project.

Relevance of Original Goals:

Despite the adjustments made in the project plan, the original goals remain largely relevant. The core objective of pushing the boundaries of accuracy, efficiency, and accessibility in indoor localization using camera images remains unchanged.

Progress and Timetable:

Progress has been steady, albeit slightly behind schedule due to the high workload from ongoing classes. The implementation phase proved to be more challenging and time-consuming than initially anticipated. As a result, adjustments were made to the project timetable to allocate more time for implementation tasks while ensuring sufficient time for post-training activities such as testing and result comparison.

Corrections and Effects:

To address the schedule deviation and ensure successful completion, corrective measures were implemented. These measures included revising the project timetable to allow for additional time for implementation tasks. By reallocating time from other project phases, such as evaluation and testing, we aimed to strike a balance between meeting deadlines and maintaining the quality of deliverables. It is believed that the effects of these changes will be positive, as the new timetable provides a more realistic and manageable timeline for project execution.

4. TASKS TO BE COMPLETED BEFORE FINAL REPORT

Before the next progress report, our focus will be on refining the conceptual design of Design A and advancing its implementation. This entails incorporating Swin Transformers [8] into the design, employing ideas from the paper An Efficient Indoor Localization Based on Deep Attention Learning Model [10], and exploring ensemble techniques with another CNN model. Transfer learning will be utilized to adapt pre-trained Swin Transformer models to the specific requirements of indoor localization. For design B, a robust object recognition model with high accuracy will be made.

Once implemented, we will rigorously test and evaluate the preliminary design of Design A. This evaluation will involve assessing the performance of Swin Transformers in capturing spatial information from input images and evaluating the impact of ensemble learning on localization accuracy. Based on the evaluation results, we will refine the detailed design of Design A, fine-tune parameters, and optimize algorithms to address any identified shortcomings or areas for improvement.

Figure [8]: Gantt chart showing tasks to be completed before the final report

5. REFERENCES

[1] Fusco, G., Coughlan, J.M. (2018). Indoor Localization Using Computer Vision and Visual-Inertial Odometry. In: Miesenberger, K., Kouroupetroglou, G. (eds) Computers Helping People with Special Needs. ICCHP 2018. Lecture Notes in Computer Science(), vol 10897. Springer, Cham. https://doi.org/10.1007/978-3-319-94274-2_13

[2] Niu, Q., Li, M., He, S., Gao, C., Gary Chan, S.-H., & Luo, X. (2019). Resource-efficient and automated image-based indoor localization. *ACM Transactions on Sensor Networks*, *15*(2), 1–31. https://doi.org/10.1145/3284555

[3] Akal, O., Mukherjee, T., Barbu, A., Paquet, J., George, K., & Pasiliao, E. (2018). A Distributed Sensing Approach for Single Platform Image-Based Localization. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. https://doi.org/10.1109/icmla.2018.00103

[4] Piasco, N. (2019). *Vision-based localization with discriminative features from heterogeneous visual data* (Doctoral dissertation, Université Bourgogne Franche-Comté).

[5] Chen, Y., Chen, R., Liu, M., Xiao, A., Wu, D., & Zhao, S. (2018). Indoor visual positioning aided by CNN-based Image retrieval: Training-free, 3D modeling-free. *Sensors*, *18*(8), 2692. https://doi.org/10.3390/s18082692

[6] Li, S., Yu, B., Jin, Y., Huang, L., Zhang, H., & Liang, X. (2021). Image-Based Indoor Localization Using Smartphone Camera. *Wireless Communications and Mobile Computing*, *2021*, 1–9. https://doi.org/10.1155/2021/3279059

[7] Li, Q., Cao, R., Liu, K., Li, Z., Zhu, J., Bao, Z., Fang, X., Li, Q., Huang, X., & Qiu, G. (2023). Structure-guided camera localization for indoor environments. ISPRS Journal of Photogrammetry and Remote Sensing, 202, 219–229. https://doi.org/10.1016/j.isprsjprs.2023.05.034

[8] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv48922.2021.00986

[9] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2016.91

[10] Abozeid, A., I. Taloba, A., M. Abd El-Aziz, R., Faiz Alwaghid, A., Salem, M., & Elhadad, A. (2023). An efficient indoor localization based on deep attention learning model. *Computer Systems Science and Engineering*, *46*(2), 2637–2650. https://doi.org/10.32604/csse.2023.037761