Navigating Indoors with Computer Vision: Exploring Deep Learning Approaches for Room-Level Indoor Localisation

Mika Senghaas (Author)
IT University of Copenhagen
jsen@itu.dk

Stella Grasshof (Supervisor)
IT University of Copenhagen
stgr@itu.dk

A Thesis presented for the Degree of
Bachelor of Science in Data Science

IT UNIVERSITY OF CPH

IT University of Copenhagen
Course Code: BIBAPRO1PE

May, 15th 2023

Contents

Abstract

In an increasingly urbanised and digitalised world, indoor localisation is becoming a necessity for a wide variety of applications, ranging from personal navigation to augmented reality. However, despite extensive research efforts, indoor localisation remains a challenging task and no single solution is widely adopted. Motivated by the success of deep learning in numerous computer vision tasks, this study explores the feasibility of deep learning for accurate room-level localisation in indoor spaces. Various neural network architectures are trained and evaluated on a novel video dataset tailored for indoor localisation. The findings reveal that deep learning approaches can provide reasonable localisation results, even when trained on a small dataset. The approach is currently limited in its ability to distinguish between visually similar and adjacent areas, as well as biases within the training data. Despite these shortcomings, the results are encouraging and inspire optimism about the method's practical viability.

1 Introduction

With the introduction of the satellite-based Global Positioning System (GPS), localisation in outdoor spaces has become more efficient and accurate than ever before. Gradual commercialisation led to the technology rapidly transforming industries and personal navigation. Today, outdoor localisation is widely considered a *solved problem*.

The same cannot be said for indoor localisation. Because the transmitted radio signals sent out by the satellites in GPS systems are not strong enough to penetrate through walls and struggle with reflections from large buildings, the technology yields inaccurate results at best, and often becomes dysfunctional in indoor spaces [14, 20].

With the ongoing urbanisation and the emergence of autonomous robots and vehicles, the need for indoor localisation technologies is growing. Over the past decade, a wide variety of solutions have been proposed. Infrastructure-based systems use radio signals transmitted by beacons, like Bluetooth [3, 5], Ultra-Wideband (UWB) [1, 2] or Wi-Fi [14, 20, 30], to localise an agent in a known environment. Infrastructure-less systems, like simultaneous localisation and mapping (SLAM) algorithms, rely solely on sensors, like cameras [6, 21, 27] or distance-measuring lasers [29] to localise an agent in an unknown environment.

While these approaches have produced remarkable results, localising agent's with centimetre accuracy, they are limited for various reasons: Infrastructure-based systems require initial setup and maintenance of the installed hardware, which makes them costly, time-intensive and difficult to implement in large environments. Current infrastructure-less systems, on the other hand, require complex hand-designed algorithms for processing the sensory information and need to be fine-tuned by experts for each indoor space, to achieve outstanding results. Much of the complexity of existing solutions is grounded in the assumption that all applications require centimetre accuracy. However, not all use-cases require such precision. For example, in a museum or shopping mall, it might be sufficient to know in which area a visitor is. In these cases, the constraint of centimetre accuracy can be relaxed, in favour of a simpler and more versatile solution.

Deep learning, which is part of a broader family of machine learning methods, has recently gained a lot of attention in the field of computer vision and proven to be a powerful tool for solving a wide variety of tasks. Common computer vision tasks are image and video classification, where the goal is to predict a label from a set of pre-defined labels for a given image or video.

Given this, it is natural to ask (a) whether indoor localisation can be phrased as a coarse-grained classification task, where labels correspond to areas in an indoor space, and (b) whether deep learning

techniques can be used to produce accurate localisation results in this setting. Therefore, this study investigates the applicability of modern deep learning techniques to indoor localisation. The main contributions can be summarised as:

1. A novel single-frame and video classification dataset tailored for indoor localisation.
2. A rigorous evaluation of several modern deep learning architectures for the task of indoor localisation, when viewed as a classification task.
3. A discussion of the results and an outlook on the applicability of a pure deep learning pipeline for indoor localisation.

2 Background

Phrasing the problem of indoor localisation as a classification problem and solving it with a pure deep learning approach requires a brief introduction to some of the fundamentals that underlie this project's methods. This section, therefore, introduces the fundamental concepts of deep learning relevant to this study.

2.1 Fundamentals of Machine and Deep Learning

Machine learning is a subfield of artificial intelligence (AI) that describes a series of techniques and algorithms that allow computers to learn from data without being explicitly programmed. One example of a machine learning task is classification, which aims to assign a discrete label $\hat{y} \in \{y_1, \ldots, y_n\}$ to an input x. The mapping from the input x to the label $\hat{y}$ is called a classifier and is often denoted as $f(x) = \hat{y}$. Typically such a classifier is trained on a large set of labelled data, called the training set, which consists of true instances x_i and their labels y_i. Using numeric optimisation algorithms, like gradient-descent, the classifier iteratively updates its parameters to minimise a loss function $\mathcal{L}(\hat{f}(x), y)$ that quantifies the error of the machine learning model.

Deep learning is a subfield of machine learning and describes a specific class of machine learning algorithms based on the theory of artificial neural networks (ANNs). ANNs are inspired by and loosely related to the structure and functioning of the neurons in the human brain. They are structured in a series of fully-connected layers, each consisting of nodes. Figure 1 shows an ANN with an input layer with three nodes, three hidden layers with seven nodes each and an output layer with three nodes. Information flows from the input layer through the hidden layers to the output layer. The information flow happens through sequential linear transformations of the input data performed by the network nodes. Specifically, each node's output is a linear transformation of the outputs of all nodes in the previous layer. Hence, it can be written as

$$z_i = \sum_{j=1}^{n} w_{ij} x_j + b_i \quad , \tag{1}$$

where x_j is the output of the j-th node in the previous layer, w_{ij} is the weight of the connection between the j-th node in the previous layer and the i-th node in the current layer, b_i is the bias of the i-th node in the current layer, and z_i is the output of the node.

Before the output z_i is passed to the next layer, it is transformed by a differentiable, non-linear activation function σ to produce an activation $a_i = \sigma(z_i)$. Once all activations in the j-th layer are computed, the next layer's activations can be computed in the same way. This process, referred to

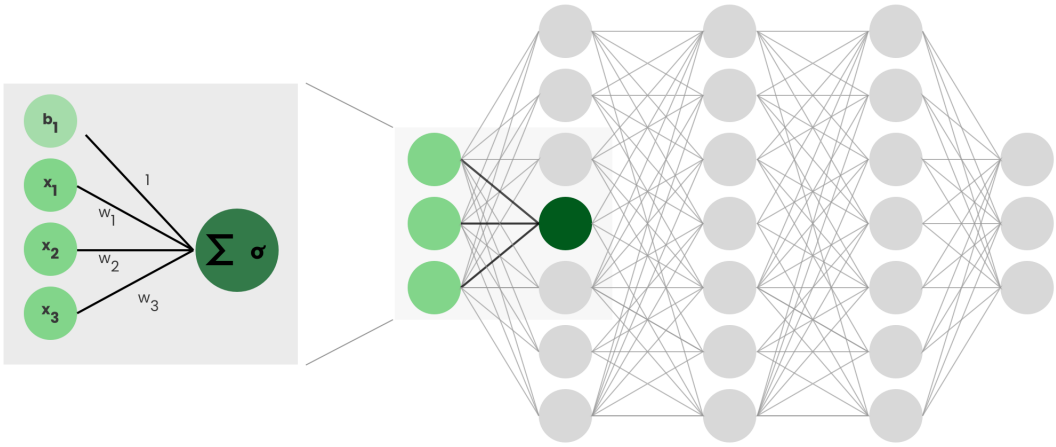


Figure 1: **Artificial Neural Network.** A schematic of the fundamental building blocks of an artificial neural network (ANN). The right Figure shows the macro structure of an exemplary ANN with an input layer, three hidden layers, and an output layer. The left Figure shows the micro structure of a single node in the network. The node performs a linear transformation of the inputs x_i and the weights w_i and adds a bias b_i. The output of the node is the result of a non-linear activation function σ applied to the linear transformation.

as forward propagation, is iteratively repeated until the activations in the output layer are computed, which are the final outputs of the network.

Critically, the linear transformations performed by each node are parametrised by weights, which are optimised during training. This allows ANNs to learn complex non-linear mappings given enough samples of the input-output relationship without explicitly programming the mapping. The generality and scalability of the method have proven ANNs to be a powerful tool for modelling complex data relationships in a wide variety of domains.

2.2 Models for Image Classification

Image classification is one of the most fundamental and widely studied tasks in computer vision and describes the process of assigning a discrete label $y \in \{y_1, \ldots, y_n\}$ to an image x. However, extracting information from images to assign a label is not straightforward because of the images' high-dimensional and unstructured nature. These characteristics make it challenging for traditional heuristic-based algorithms to extract meaningful information, which has long limited the capabilities of computer vision systems. However, this has changed with the advent of deep learning and the introduction of a particular type of neural network called convolutional neural network (CNN).

CNNs are a type of neural network specifically designed to process visual information. Traditionally, they are organised hierarchically and consist of convolutional, pooling and fully connected layers, as shown in Figure 2. Convolutional layers are the core of CNNs and are responsible for extracting features from the input. Each convolutional layer consists of a set of filters, where each filter, sometimes called kernel k, is a three-dimensional matrix of weights with dimensions $c_i \times h_k \times w_k$, where c_i is the number of channels in the input, and h_k and w_k are the height and width of the filter.

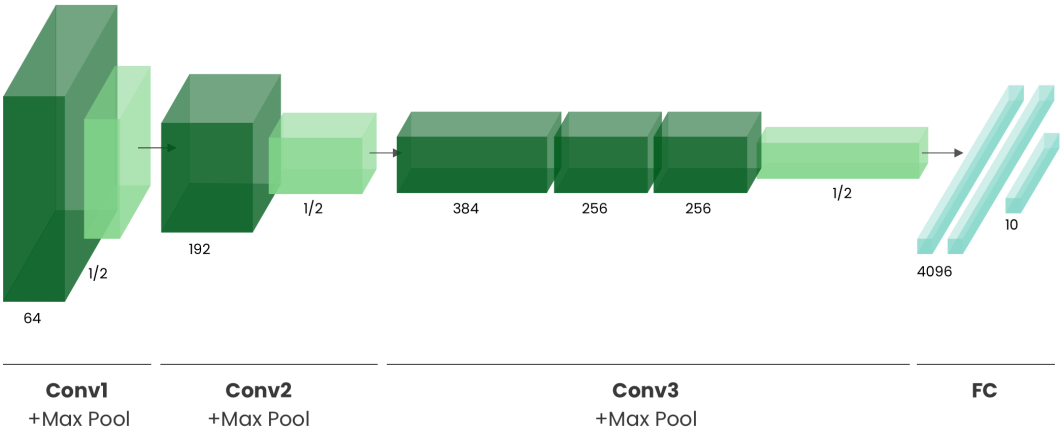


Figure 2: **Convolutional Neural Network.** A schematic of a traditional convolutional neural network (CNN), designed for image classification (here: AlexNet [22]). The network consists of a series of 2D convolutional layers, pooling layers and fully-connected layers, whose outputs are displayed as dark-green, light-green and blue boxes, respectively.

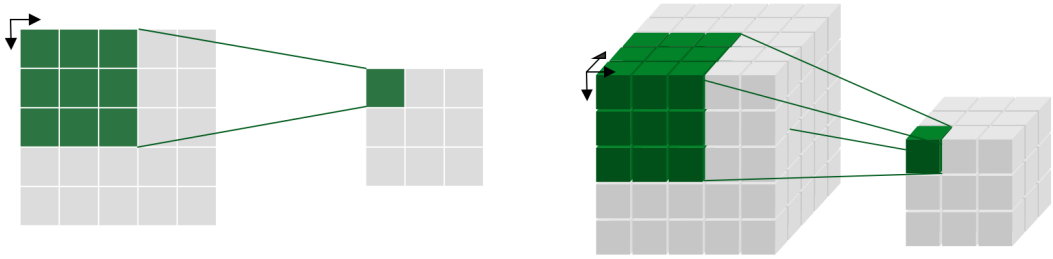
Given an input x with dimensions $c_i \times h_i \times w_i$, a single convolutional filter k produces a feature map z by sliding the filter across the spatial dimensions of the input, as shown in Figure 3a, and computing the convolution at each position i, j (Equation: 2).

$$z_{i,j} = \sum_{c=1}^{c_i} \sum_{m=1}^{h_k} \sum_{n=1}^{w_k} k_{c,m,n} x_{c,i+m,j+n} \tag{2}$$

The output z is a two-dimensional matrix called feature map. Within a single convolutional layer, multiple filters are applied to the input, which results in multiple feature maps. After applying a non-linear activation to each feature map, they are stacked along the channel dimension to form the output of the convolutional layer, which is passed to the next layer. Pooling layers are often interleaved with convolutional layers to reduce the dimensionality of feature maps. The pooling operation is similar to convolution, as a kernel is slid over the spatial dimensions of the input. However, instead of computing a weighted sum, the pooling operation computes a statistic, such as the maximum (Max Pooling) or average (Average Pooling), of the values in the kernel.

CNNs have been found to work specifically well for data with a spatial structure, such as images because convolutions model the inherent nature of images: While a stand-alone pixel is not as informative, the value of a pixel in the context of its neighbouring pixels is. The convolution operation naturally captures this characteristic. Furthermore, sliding filters across the input, can capture the same feature independently of its location in the image, making CNNs invariant to translation. This is a desirable property for image classification, as the location of features in an image may not be relevant for the final classification.

The first CNN-based architectures date back to the 1960s and have succeeded in simple image classification tasks [23]. However, it was not until 2012, when the CNN-based architecture AlexNet [22] won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [7], that CNNs arrived in the mainstream of computer vision. Since then CNNs have been proposed in various forms, mainly differing in the network's structure and depth and the width and resolution of the filters. To this day,



(a) 2D Convolution (b) 3D Convolution

Figure 3: **2D and 3D Convolution.** A simplified illustration striding of 2D and 3D convolutional filters across an input. For visualisation purposes, a single channel is shown. A 2D convolutional filter (a) slides across the spatial dimension (height and width) of an input. A 3D convolutional filter (b) slides across the spatiotemporal dimension (height, width, and time) of an input.

CNNs still rank amongst the top-performing methods in image classification benchmarks.

2.3 Models for Video Classification

Video classification can be seen as a generalisation of image classification: Instead of assigning a label to a single image, video classifiers assign a label y or a sequence of labels $y_1, \ldots, y_t$ to a video sequence $x = (x_1, \ldots, x_T)$, where x_t is a frame of the video.

The difference is subtle yet essential because it introduces a temporal dimension. It is generally assumed that the temporal dimension is critical for video classification because it provides additional information about the video. For example, it might be challenging to determine whether a person is running or walking from a single frame but trivial given the motion captured by a sequence of frames.

Motivated by the need for a powerful model to capture the semantic content of video data and the success of CNNs in image classification, researchers have started to apply CNNs to video classification [18, 4, 33, 10, 11], where the networks have access to the complex spatiotemporal evolution of the video. Over the course of the last decade, various approaches based on CNNs have been proposed that exhibit different connectivity patterns for modelling the temporal dimension of the video [18].

A naive solution ignores the temporal dimension entirely and predicts a label for each frame and then averages the predictions. Despite the simplicity of the approach, such models have been shown to perform surprisingly well [18]. Different aggregation methods have been proposed, such as averaging [18], majority votes, or using a neural architecture, such as recurrent neural networks (RNNs), to learn the importance of each frame [8] for the final prediction.

CNN-based architectures that model the temporal dimension directly usually leverage 3D convolutions [33, 4]. 3D convolutions are a natural extension of 2D convolutions, as defined in Section 2.2. Instead of sliding a convolutional filter across the spatial dimension of the input, as shown in Figure 3a, a 3D-convolutional filter slides across the spatiotemporal dimension of the input , illustrated in Figure 3b), which produces feature maps in spatiotemporal space that are learned jointly. 3D convolutions have been used in many different variants. Architectures range from pure 3D convolution networks [4, 33] to hybrid architectures that combine 2D and 3D convolutions [10, 11]. Overall, it

Figure 4: **Raw Frames.** An example of the high-resolution (2426×1125) raw frames captured by the mobile device. The Figure shows five frames that were extracted at a sampling rate of 3 frames per second (FPS) from the original video sequence. The location is the green area on the first floor (First Floor Green Area).

can be summarised that, despite the introduction of large-scale datasets for video classification [19], the research field has not agreed on the best approach for video classification.

3 Data

To train a deep learning model for indoor localisation, a labelled dataset is required for training and evaluation. The following section describes the collection, annotation and pre-processing of a novel video dataset tailored for indoor localisation. Because the study investigates models trained on static frames and sequences of frames, the pre-processing of the data differs and is therefore described separately.

3.1 Data Collection

The raw video data was collected from a single mobile device camera at a frame rate of 30 FPS in high resolution (2426×1125). This process yielded a set of $n = 53$ videos $V = \{v_1, ..., v_n\}$, where each video v_i is a sequence of k frames $f_1, ..., f_k$, and each frame is f_i is an RGB image with dimensions $3 \times 2426 \times 1125$. The number of frames per video differs, with an average of $\sim$1700 frames per video, or 57 seconds.

The mobile device was hand-held by a human agent while walking around the main building of the Southern Campus of the Copenhagen University (Danish: Københavns Universitet, KU) in Copenhagen S, Denmark. The location was deemed compatible with this study as it showcases distinctive learnable indoor features (e.g. coloured walls, characteristic architectural structures) but also challenges the model, for example, due to areas that are visually similar to each other (like libraries and corridors).

Five exemplary frames from a single raw video are shown in Figure 4 for illustration purposes.

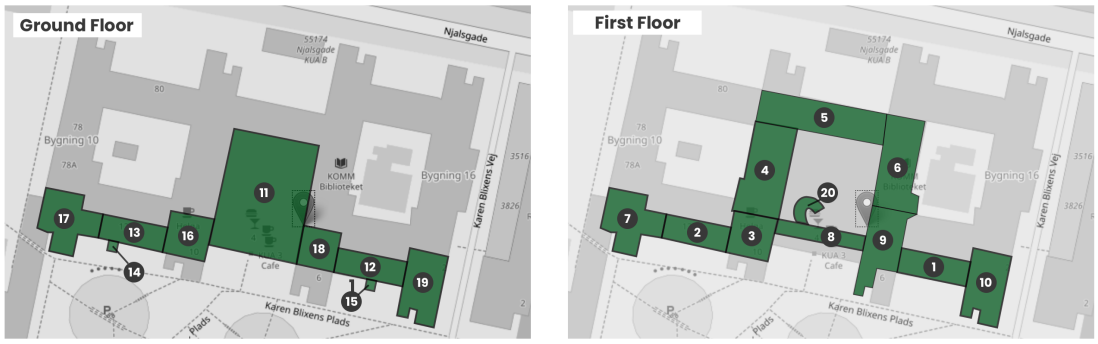


Figure 5: **Location Labels.** The Figure shows the floor plan for the ground floor (left) and first floor (right) of the main building of the Southern Campus of the Copenhagen University. The 20 location labels considered in this study are numbered and mapped to a coloured region on the floor plan.

3.2 Data Annotation

Each video v_i is associated with a set of location labels $L = \{l_1, ..., l_n\}$, where each location label l_i identifies an agent's location at a specific time in the video. For the scope of this project, a subset of the areas on the first two floors of the main building was considered and grouped into 20 different location areas. Each area represents a class in the classification task and is identified by a descriptive name and integer. Figure 5 shows the floor plan of the first two floors of the building, where each considered location class is mapped to a coloured region. The location labels were assigned in close correspondence to the original floor plan. However, this led to some classes being a lot larger than others, resulting in a class imbalance. Annotation was performed manually by a single human agent. Because changes in the location labels only occur at the transition of rooms, the annotation process was simplified by annotating the starting and ending time stamps of a location label, which were later pre-processed to frame-by-frame annotations.

3.3 Data Splits

Out of the total 53 videos that were recorded, 37 were used for training, and 16 were used for testing. Notably, the videos in the training split were recorded in a single session. In contrast, the videos in the test split were recorded on four separate days, two to four weeks after the training data had been recorded. This ensured that the models were tested against unseen data to assess their generalisation capabilities more accurately. Indeed, the test data was recorded in different weather conditions, at different times of the day, with different lighting conditions. By pure chance, one of the areas was repainted during the time between the recording of the training and test data. With all these changes in mind, the test data is expected to be as different from the training data as it would be in a real-world scenario.

3.4 Data Processing

Both single-frame and video classification models expect an input-location pair for training. For single-frame models, an input is a frame tensor f_i with dimension $3 \times h \times w$. For video classification models, an input is a clip tensor c_i with dimension $s_v \times 3 \times h \times w$, where s_v is the number of frames in a clip. Both models expect a fixed input size and each frame to be standardised. While the standardisation and spatial downsampling of the frames is almost equivalent for both model groups, the temporal downsampling procedure had to be designed carefully to allow for proper comparison between the two model groups. Therefore, this section first explains the temporal downsampling of the frames for the two model groups individually and then jointly describes the spatial downsampling, standardisation, and location matching procedure.

Temporal Downsampling. A video classification model expects a single clip c_i as input, which is a fixed-sized sequence of s_v frames sampled from a video v_i at a sampling rate r_v. A sampling rate is related but not equivalent to frames per second (FPS). A sampling rate of $r_v = 5$ means that every fifth frame from a video v_i was extracted, equivalent to $30/5 = 6$ FPS. Given this, the video dataset is constructed by extracting uniformly sampled clips from the videos. For example, the first clip c_1 is extracted from the frames $[f_1, f_{1+r_v}, f_{1+2r_v}, ..., f_{1+(s_v-1)r_v}]$. The second clip, c_2, follows the same pattern, and starts at frame $f_{1+(s_v-1)r_v+1}$, the immediate successor of the last frame of the first clip. Within this study, clips were extracted in a way that ensured that the location label l_i is the same for all frames $f_i \in c_i$.

For single-frame classifiers, all frames in a video v_i could be used as input. However, because of the strong local correlation between adjacent frames, it was hypothesised that models would overfit the training data. For this reason, and in an attempt to assimilate the single-frame and video datasets, a fixed sampling rate of $r_f = 5$ was used to downsample the temporal dimension of the training data split. However, a sampling rate was not adopted for the test split to test the model on the full set of frames.

Figure 6 illustrates the temporal downsampling process for both the single-frame and video datasets on a small, illustrative example and highlights that because of the difference in sampling rates, the extracted frames can differ between the two datasets.

Spatial Downsampling. The spatial downsampling of the frames was done by resizing the frames to a fixed size of $h \times w$ pixels. The height and width of the input frames are tied to the model architecture and are individual to each model. However, the input frames are typically square, with a side length of between $h = w = 182$ or $h = w = 224$ pixels. Resizing was performed non-aspect-preserving to compress as much of the original viewport as possible into the resized frame.

Standardisation. Finally, the individual frames were standardised. An RGB image is a three-dimensional tensor where each element represents the intensity of a single colour channel at a single pixel location in the range $[0, 255]$. Because neural networks are sensitive to the scale of the input data, the intensity values of the frames were first scaled to the range $[0, 1]$ and then normalised to have a mean of 0 and a standard deviation of 1. The standardisation was performed on a per-channel basis, meaning that a colour channel x_c of a frame x was standardised as,

$$x'_c = \frac{x_c - \mu_c}{\sigma_c}, \quad c \in \{R, G, B\}, \tag{3}$$

where x'_c and x_c represent all pixels of the colour channel c of the standardised and original frame x, respectively. μ_c and σ_c are the mean and standard deviation of the colour channel c across all pixels from the dataset. Because all models were fine-tuned, the mean and standard deviation were not

Extracted Frames in **Single-Frame Dataset**

Extracted Frames in Clips in **Video Dataset**

Figure 6: **Temporal Downsampling.** An illustration of the temporal downsampling process for the single-frame and video datasets to illustrate the concept of sampling rates (r_f/r_v). The same selection of frames as in Figure 4 is used. For the single-frame dataset, frames are sampled at a rate of $r_f = 2$ (every other frame), resulting in a total of $n_f = 3$ frames. For the video dataset, frames are sampled at a rate of $r_v = 3$ (every third frame), and a clip (illustrated by a surrounding light-green box) consists of $s_v = 2$ frames. Frames that are not sampled into the dataset are visualised in greyscale.

computed from the video location dataset but instead taken from the respective pre-training dataset, ImageNet [7] and Kinetics [19] for the single-frame and video datasets, respectively.

Location Matching. Associating a frame f_i or a clip c_i with a location label l_i was trivial. Because the timestamp of extraction of each frame and the location label is known, the location label l_i of a frame f_i is the location label present at the time of extraction of the frame. Similarly, because clips do not overlap multiple location labels, a clip c_i can safely be associated with the location label l_i present at the time of extraction of the clip's first frame.

All of the above steps yielded the two datasets D_F and D_V, for the single-frame and video classification tasks, respectively.

4 Methodology

The study follows a standardised methodology to assess the capabilities of different deep learning models. As seen, first a self-gathered dataset is collected, annotated and preprocessed according to the specification of the problem setting. Next, a series of deep learning models are selected and trained under constant conditions on the same dataset and evaluated using commonly used metrics for performance and efficiency. Finally, the results are analysed to discuss the applicability of deep learning models for indoor localisation.

One major distinction between the different models is whether or not they operate on a single frame or a sequence of frames. Because this distinction impacts all steps of the study, from data processing to model evaluation, the following two different problem settings are distinguished throughout the entire study:

1. **Single-frame classification**: Given a continuous stream of frames, the task is to classify each frame individually.
2. **Video classification**: Given a continuous stream of frames, the task is to classify fixed-sized clips of the video.

4.1 Models

Twelve different models were trained and evaluated in this work. Following the distinction made in Section 4, the models are grouped into single-frame and video classification models.

Table 1 gives a comprehensive overview of all models. The table shows the release year, the sampling rate r_f for single-frame classification models and the sampling rate r_v and clip size s_v for video classification models. The table also shows the input size, the number of parameters and the number of floating point operations (FLOPs) for a single forward pass. Finally, the table shows the Top-1 accuracy on the ImageNet dataset [7] for single-frame classification models and the Top-1 accuracy on the Kinetics dataset [19] for video classification as an indicator of the model's performance on a generic classification task.

The following section briefly describes the main architectural features of each model.

4.1.1 Single-Frame Classifiers

Alexnet [22] is a convolutional neural network introduced by Krizhevsky *et al.* in 2012. It was the first deep neural network to win the ImageNet Large Scale Visual Recognition Challenge [7] and is considered one of the first successful applications of deep learning to image classification.

Table 1: **Model Overview.** The Table shows all models that were evaluated in this work. The models are split into two categories: single-frame models and video models. For each model, the Table reports the release year (Release), the frame rate (Rate) of the training data, the number of frames per clip (F/C; *only applicable to video classifiers*), the spatial resolution (Size) of the input images, the number of parameters in millions (Params), the number of floating point operations in billions (FLOPs) and the benchmark Top-1 accuracy (Acc1) on ImageNet [7] for single-frame classification models and the Top-1 accuracy on the Kinetics [19] dataset for video classification models. The Table is sorted by release date within each group.

	Model	**Release** (Y)	**Rate** (r_f/r_v)	**F/C** (s_v)	**Size** (h, w)	**Params** (M)	**FLOPs** (G)	**Acc@1** (%)
Single-Frame	AlexNet [22]	2012	5	-	224	61.1	0.71	56.52
	ResNet18 [13]	2015	5	-	224	11.7	1.81	69.76
	ResNet50 [13]	2015	5	-	224	25.6	4.09	76.13
	DenseNet 121 [16]	2016	5	-	224	7.0	2.88	74.43
	MobileNet V3 [15]	2019	5	-	224	3.5	0.32	71.88
	ViT-B-16 [9]	2020	5	-	224	86.7	17.56	81.07
	EfficientNet V2 S [31]	2021	5	-	224	21.5	8.37	84.23
	ConvNext Tiny [25]	2022	5	-	224	28.2	4.46	82.52
Video	R(2+1)D [34]	2018	4	16	182	28.11	76.45	76.01
	Slow R50 [11]	2018	8	8	224	32.45	54.52	74.58
	SlowFast R50 [11]	2018	8	8	224	34.57	65.71	76.94
	X3D S [10]	2020	6	13	182	3.5	2.96	73.33

Architecturally, it consists of five convolutional layers with occasional max-pooling layers in between, followed by three fully connected layers, as illustrated in Figure 7.

After the success of AlexNet, the speed of research in deep neural networks significantly picked up. Networks with increasing depth, such as VGG [28], were found to perform better but led to a new research problem: As information about the input or gradient passes through many layers, it can vanish during gradient-descent based learning, leading to stagnation during training. The phenomenon of "vanishing gradients" was visible in larger training errors for deeper networks and coined the "degradation" problem [13].

The **ResNet** [13] architecture is considered a cornerstone of modern deep learning history as it introduced the concept of residual connections. In a residual subnetwork of L layers H_l, the output x_l is computed as the sum of the input to the subnetwork x_{l-1} and the output of the subnetwork $H_l(x_{l-1})$, i.e. $x_l = x_{l-1} + H_l(x_{l-1})$. The introduction of this identify mapping to the output of the subnetwork was shown to facilitate signal propagation in forward and backward paths, and, in connection with batch normalisation [17], allowed to train networks of unprecedented length. The original paper introduced several architectures with different depths, but in this work, only ResNet18 and ResNet50, with 18 and 50 layers, respectively, are considered.

DenseNets [16] extend the idea of residual connections. At the core of the DenseNet architecture is the concept of "dense blocks": Similar to a residual block, it is a subnetwork of L layers denoted as H_l. However, unlike the residual block, each layer in the dense block receives the concatenation of all preceding layer outputs as an input, such that $x_l = H_l([x_0, x_1, \ldots, x_{l-1}])$. Dense blocks are compute-intensive, but the strong connectivity between layers supports feature reuse, which allows the architecture to be shallower than its predecessors. Again, the authors introduced several architectures

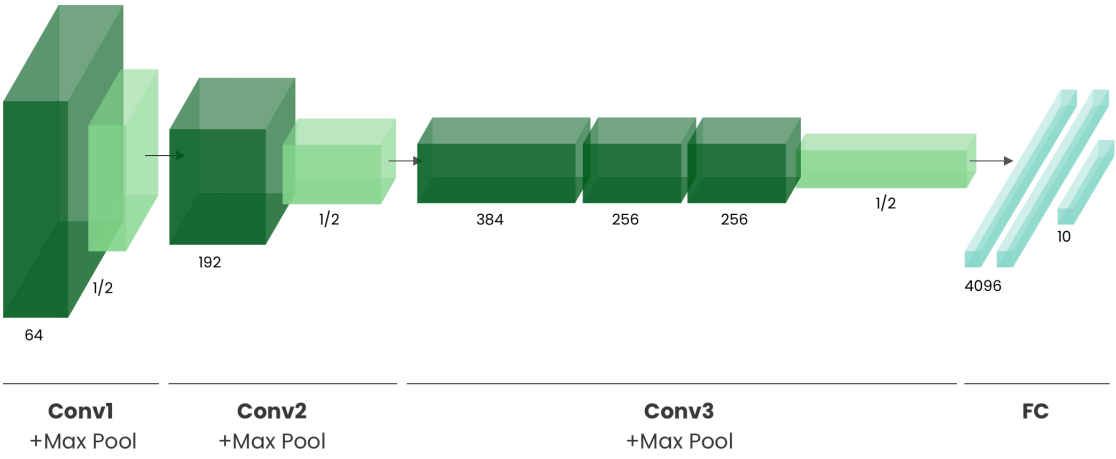


Figure 7: **AlexNet Architecture.** An illustration of the architecture of AlexNet [22]. Each box represents the output of a layer. Dark-green boxes are outputs of convolutional layers, light-green boxes are outputs of max-pooling layers and blue boxes are outputs of fully connected layers. The number of channels is denoted below each convolutional layer. Spatial downsampling is only performed by the first convolutional layer and all pooling layers. The downsampling factor is denoted below the pooling boxes.

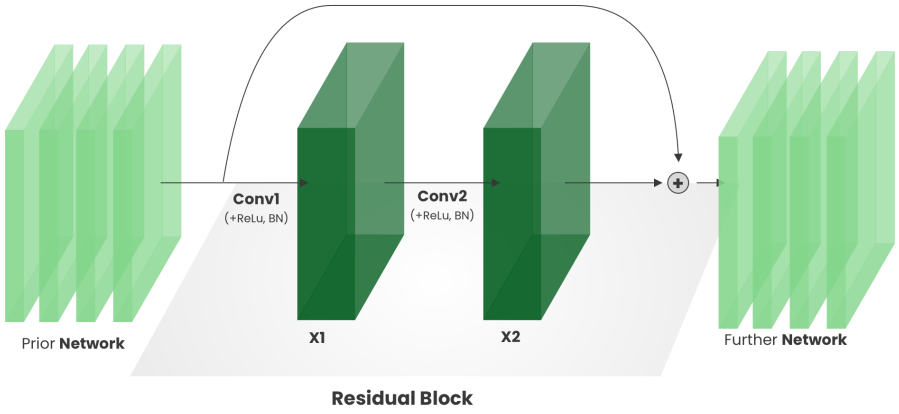


Figure 8: **Residual Connection.** An illustration of a generic residual block, as defined by He *et al.* [13]. A subnetwork of $L = 2$ layers (consisting of convolution, batch normalization and ReLU layers) is applied to an input x_0. The first layer produces an output x_1, and the second layer an output x_2. The output of the residual block is given by $x_2 + x_0$ through the skip connection.

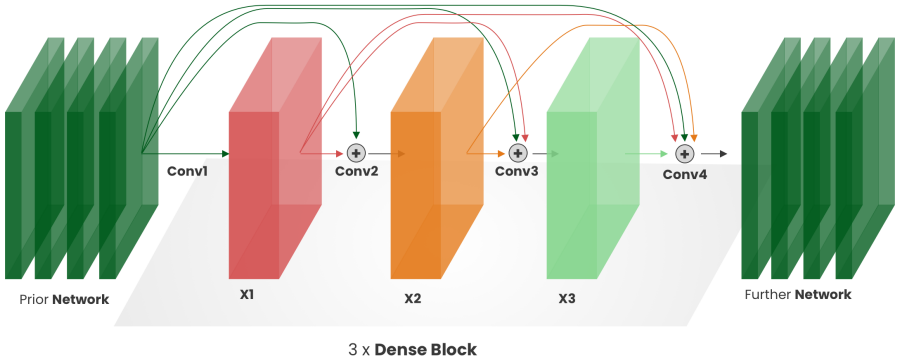


Figure 9: **Dense Block.** An illustration of a generic dense block, as defined by Huang *et al.* [16]. A subnetwork of $L = 4$ layers (consisting of convolution, batch normalization and ReLU layers) is applied to an input x_0. The i-th layer produces an output x_i. The input to the i-th layer is the concatenation of the output of all previous layers, so $H_i(x_i) = H_i([x_0, x_1, \ldots, x_{i-1}])$. The final output of the denseblock is the concatenation of the output of all layers, so $x_4 = [x_0, x_1, x_2, x_3]$.

with different depths. Here, only DenseNet 121 is considered.

MobileNet V3 [15] is the third iteration of the MobileNet architecture, which was introduced by Howard *et al.* in 2019. The main contribution of MobileNets is the introduction of depth-wise separable convolutions. A depth-wise separable convolution factorises a regular 2D convolution into two separate steps: First, a single 2D filter is applied to each input channel separately, and then a 1x1 convolution is applied to the output of the previous step. This architectural change makes MobileNets significantly more efficient with only minor performance losses, which made them popular for application on low-compute devices like mobile phones. MobileNet V3 is the most recent iteration of the architecture, and its smallest variant, MobileNet V3 Small, is used in this work.

EfficientNet V2 [31], proposed by Tan *et al.* in 2021, is searching for a compute-optimal CNN architecture. The paper's main finding is a scaling law, which states that for a given baseline network, e.g. scaling up a network's depth, width and resolution with a constant factor leads to improved performance and efficiency. In the original paper, the authors scale up previous state-of-the-art models like ResNet and MobileNet, which showcase state-of-the-art performance and make up a new family of models called EfficientNets. Here, EfficientNet V2 S is used, which is the smallest variant of the EfficientNet V2 family.

With the ground-breaking paper "Attention is all you need" [35], Vaswani *et al.* introduced the Transformer architecture in 2017. Although initially designed for machine translation, the architecture was quickly adapted to other tasks in natural language processing, superseding previous state-of-the-art models on various benchmarks.

In 2020, **Vision Transformer** [9] was introduced as one of the first Transformer-based models adapted for computer vision tasks. The main contribution of the model is the patch embedding mechanism, which overcomes the architectural mismatch between the two-dimensional input of images and the one-dimensional input of Transformer architectures. Dosovitskiy *et al.* propose to split a two-dimensional image into a sequence of fixed-sized patches, which are then flattened and embedded into a sequence of "tokens". Together with positional encodings for the position of the patch in the original image, this sequence of tokens can be fed into a regular Transformer architecture. The au-

thors introduce several configurations of the architecture. In this work, the base Vision Transformer, ViT-B-16, which is among the models' smallest variants, is used.

Since then, many variants of the Vision Transformer have been proposed. They either decrease the computational complexity of the architecture [32, 12] or improve the performance by recovering some of the inherent architectural advantages of convolutional neural networks [24]. However, these architectures are not considered in this work.

Finally, **ConvNext** [25] is one of the most modern convolutional neural network architectures considered in this work and was proposed by Liu *et al.* [25] in 2021. Against the trend of Transformer-based architectures in computer vision tasks, ConvNext is a pure convolutional architecture. ConvNext uses the ResNet 50 [13] architecture as a baseline and performs gradual modernisation steps, ranging from different optimisation algorithms and filter dimensions to network depth and width. The authors show that the resulting convolutional architecture is competitive with the Transformer-based architectures. Within this study, ConvNext Tiny, which is the smallest variant of the ConvNext family, is used.

4.1.2 Video Classifiers

Early attempts of using 3D convolutional filters for video classification tasks [4, 33] show potential for modelling the temporal dimension of video data. However, the authors also show that simple single-frame models perform surprisingly well and that the temporal dimension only provides marginal gains. This finding motivated Tran et. al to investigate different architectural designs for video classification tasks. They contrast regular 3D convolution, mixed 2D and 3D convolution, 3D convolution and 3D convolution with factorised filters, which they coin **R(2+1)D** [34]. They find that the R(2+1)D architecture improves the performance in benchmarks, such as the Kinetics dataset, by a noticeable margin.

In the same year, Feichtenhofer *et al.* [11] proposed **SlowFast** [11], a two-stream architecture for video classification tasks. SlowFast has two separate pathways for processing video data. The first pathway, the slow pathway, processes the video data at a low frame rate, which allows it to capture the spatial information of the video data. The second pathway, the fast pathway, processes the video data at a high frame rate, which allows it to capture the temporal information of the video data. The authors show that combining the two pathways improves performance over single-stream architectures. This study considers both the single-stream slow architecture **Slow R50** and the two-stream architecture **SlowFast R50**.

Finally, the **X3D** architecture reviews the design choices of previously proposed architectures. The authors find that previous approaches differ mainly by varying the temporal dimension (frame rate and number of frames) and the spatial dimension (resolution, depth and width of filters). Jointly studying the effects of scaling the temporal and spatial dimensions, the authors propose a family of architectures that gradually scales up the network's depth, width and resolution , while keeping the computational complexity constant. In this work, the smallest variant of the X3D family, X3D S, is used.

4.2 Training

All models were implemented using the deep learning framework `PyTorch`. The model architectures were taken from the `Torchvision` and `Pytorchvideo` libraries and initialised with the default pre-trained weights provided by the libraries. Single-frame and video classifiers were pre-trained on the ImageNet [7] and Kinetics [19] datasets, respectively. Fine-tuning was performed remotely on the high-performance cluster (HPC; Appendix 7.2) of the IT University of Copenhagen using GPU-accelerated

training.

All models were optimised to minimise the cross-entropy loss function $\mathcal{L}$ as

$$\mathcal{L}(\hat{y}, y) = -\sum_{i=1}^{L} y_i \log(\hat{y}_i) \quad , \tag{4}$$

where $\hat{y}$ is the predicted probability distribution over the L location labels, and y is the one-hot encoded ground truth location label for a single frame or video. The function penalises the model for predicting a low probability for the ground truth label. For the models to learn the relationship between frames/ clips and location labels, the gradient of the loss function with respect to all model parameters is computed and AdamW [26], the de facto standard optimiser for deep learning models, is employed to update the model parameters at a constant learning rate of 10^{-4}. No learning rate warm-up or scheduling was performed during training.

Mini-batch training was used for all models with a batch size of 16 for single-frame models and 4 for video classification models due to memory constraints. Because of the high computational complexity of 3D convolutions, the video classification models take longer to converge and were therefore trained for 15 epochs, while the single-frame models were trained for 10 epochs.

4.3 Evaluation

A series of performance and efficiency metrics are computed to assess all models in terms of their performance and efficiency. All metrics are computed on the test split, separated from the original dataset before training, as described in Section 3.

Quantifying Performance. In the context of indoor location classification, a model is considered to perform well if it is able to accurately predict the location given a single frame or video clip. An intuitive way to quantify the performance of a model is to compute the Top-K multi-class accuracy:

$$\text{Top-K Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y_i \in \mathcal{Y}_i) \quad , \tag{5}$$

Here, N is the number of samples, y_i is the ground truth label for sample i and $\mathcal{Y}_i$ is the set of the Top-k predictions for sample i. The indicator function $\mathbb{I}$ is one if the ground truth label is in the set of Top-k predictions and 0 otherwise. Within the context of this report, the Top-1 and Top-3 accuracy are computed. The Top-1 accuracy is most relevant as the top prediction is typically used in applications. The Top-3 accuracy is also computed to better understand the cases where the model fails to predict the true location as the top prediction.

Furthermore, the Macro F1-score is computed. The F1 score (Equation 8) is a class-specific metric that measures the harmonic mean of precision and recall. The precision P_i of a classifier towards some class i can be interpreted as the probability that a sample classified as class i is actually from class i, and can be written as

$$P_i = \frac{TP_i}{TP_i + FP_i} \quad . \tag{6}$$

The recall R_i is the probability that a sample from class i is classified as class i, and can be written as

$$R_i = \frac{TP_i}{TP_i + FN_i} \quad . \tag{7}$$

High precision and recall values are desirable, so an ideal classifier would have high precision and recall. The F1-score is a way to combine both metrics into a single metric by computing their harmonic mean, as

$$F1_i = \frac{2 \cdot P_i \cdot R_i}{P_i + R_i} \quad . \tag{8}$$

Finally, the Macro F1-score (Equation 9) is computed by averaging the F1-scores for each location label, as

$$\text{Macro F1-score} = \frac{1}{L} \sum_{i=1}^{L} F1_i \quad . \tag{9}$$

Here, each location label is weighted equally, regardless of the number of samples that belong to the label. For the scope of this project, the metric was found to be a good extension to the Top-K accuracy, as it gives insights into potential class imbalance issues.

Quantifying Efficiency. Efficiency is critical for real-time inference, especially on mobile devices. For this reason, direct proxies for a model's efficiency are computed using the `PyTorch Benchmark` library, which allows tracking various metrics. Within this project, the number of floating point operations (FLOPs), the mean inference time per sample (latency), and the mean number of samples per second (throughput) are considered. For real-time frame classification, a throughput of at least 30 frames per second is desirable, while for video classification, a lower throughput can be accepted.

All benchmarks were computed on a desktop CPU. While the compute resources are likely to be different on a mobile device, the results still give a good indication of the relative efficiency of the models and allow to extrapolate the insights to the performance on a mobile device. It is to be noted that latency and throughput are inversely proportional to each other: Less inference time per sample (low latency) leads to a higher number of inferences per second (high throughput). However, as inference times vary significantly between models, it is helpful to compute and visualise both metrics to accentuate high or low-throughput models.

Understanding Model Behaviour. To understand the model behaviour in more detail, the confusion matrix and a subset of the misclassified samples were manually inspected for the best-performing single-frame and video classification model.

A confusion matrix is an $L \times L$ matrix, where L is the number of location labels. The entry at index (i, j) is the number of samples that were predicted to belong to location label j, given the ground truth label i. Confusion matrices are a traditional tool in classification tasks, as they give a good overview of a model's performance on the different classes and can highlight regularly confused classes visually.

Given the insights from the confusion matrix, a subset of the misclassified samples was manually inspected to understand why the model failed to predict the correct location label. This analysis should unveil what situations are particularly challenging for the models and highlight potential areas for improvement.

Table 2: **Results.** The Table shows the performance and efficiency metrics for all trained models. The models are grouped by their type (single-frame or video). The performance metrics are the Top-1 accuracy (Acc@1), Top-3 accuracy (Acc@3) and Macro F1-Score (Ma.-F1). The efficiency metrics are the number of floating point operations (FLOPs) per inference, the mean inference time in milliseconds per prediction (Latency) and the mean number of predictions per second (Throughput). The best-performing and worst-performing model in each category is highlighted in **green** and **red**, respectively. The metrics are computed on the test split of the respective dataset. *SlowFast R50 could not be benchmarked because of limitations of the `PyTorch Benchmark` library.*

	Model	Acc@1 (%)	Acc@3 (%)	Ma.-F1 (%)	FLOPs (G)	Latency (ms/Pred)	Throughput (Preds/s)
Single Frame	AlexNet	75.00	89.28	74.74	0.71	17.13	58.43
	ResNet18	79.91	93.38	77.54	1.82	32.794	30.49
	ResNet50	82.54	93.75	79.34	4.12	56.149	17.81
	DenseNet121	78.21	91.62	77.10	2.88	56.022	17.87
	MobileNet V3 Small	78.06	91.11	76.88	**0.06**	**9.168**	**109.07**
	EfficientNet V2 Small	80.92	94.55	77.54	2.88	50.795	19.69
	ViT B-16	78.29	93.18	77.76	17.59	125.113	7.99
	ConvNext Tiny	**83.59**	94.50	**79.78**	4.47	49.345	20.27
Video	R(2+1)D	80.78	**97.15**	73.66	**93.72**	**1237.00**	**0.81**
	Slow R50	77.46	94.80	69.97	42.00	791.54	1.26
	SlowFast R50	77.46	91.33	73.32	-	-	-
	X3D	**58.72**	**68.68**	**31.43**	2.85	336.81	2.97

Finally, the best-performing single-frame and video classification models were used to continuously predict the location label of a subset of the raw videos in the test split to mimic a real-world deployment scenario. Again, the analysis of the results should drive focus to potential areas of improvement.

5 Results

When carefully designed and trained, computer visions models are capable of providing reasonably accurate predictions on indoor localisation when phrased as a classification task. The detailed results of evaluating the performance and efficiency of the different models are presented in Table 2 and discussed in the following.

5.1 Performance Analysis

Surprisingly, even simple single-frame classification models are capable of providing a reasonable solution to the task of indoor localisation. Despite the lack of information about the temporal context of the frames, the best-performing single-frame classifier (ConvNext Tiny) achieves a Top-1 accuracy of 83.59% and a Top-3 accuracy of 94.50%. This is a significant finding as it proves that, in most cases, the static information of the frames is sufficient to determine location in indoor spaces accurately.

The choice of the model architecture affects the overall performance, but only to a small degree. The Top-1 accuracy of all single-frame classifiers ranges between 75% and 84% - only a difference of 9 percentage points between the worst-performing model (AlexNet) and the best-performing model

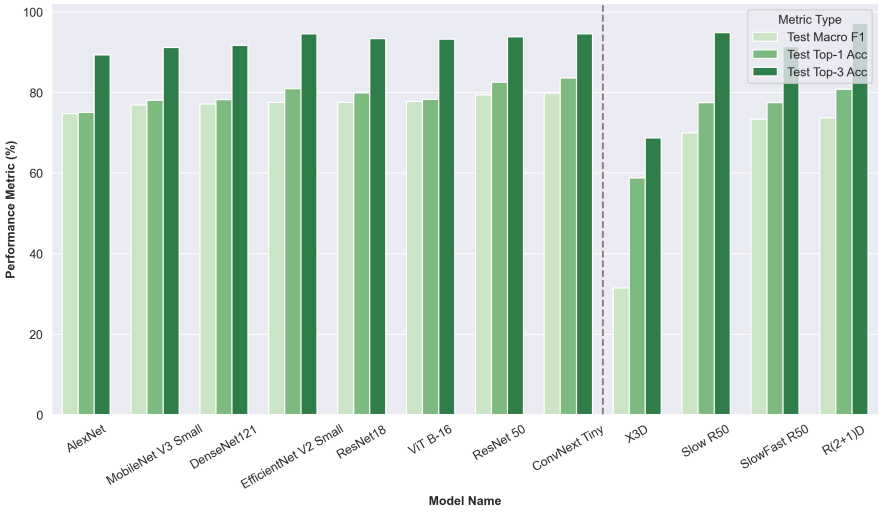


Figure 10: **Performance Metrics.** The Figure shows the performance metrics, Macro F1, Top-1 Accuracy and Top-3 Accuracy, for all trained models on the test split. A grey, dotted line separates the single-frame classifiers (left) from the video classifiers (right).

(ConvNext Tiny). Similarly, the Top-3 accuracy ranges between 89% and 95%. Generally, the results follow the performance that is expected given the ImageNet benchmarking results (Table 1). ResNet50 outperforms ResNet18, which outperforms AlexNet. The most modern model considered in this study, ConvNext Tiny, performs best overall. The only exception is ViT B-16, which performs worse than expected, given its dominance in most computer vision benchmarks in recent years. It is likely that given its significantly larger size, it struggles with undertraining. The Macro F1-Score was observed to be lower for all models but only marginally. This suggests that the models are generally not affected negatively by the class imbalance of the dataset.

It is generally assumed that knowledge about the temporal context of the video is beneficial for video classification tasks, as it allows the model to understand the motion of objects in the scene. For example, in the context of indoor localisation, the temporal context of the video could be helpful when a subset of frames is occluded by a person walking through the scene, the camera transitions between rooms, or when the camera is moved very suddenly.

However, the results suggest that video classifiers do not improve the Top-1 accuracy. The best-performing video classifier, R(2+1)D, achieves a Top-1 accuracy of 80.78%, almost 3 percentage points lower than the best performing single-frame classifier, ConvNext Tiny. The same goes for the Slow R50 Family, which achieve a Top-1 accuracy of 77.46%. The X3D model is an outlier in all models, achieving a Top-1 accuracy of only 58.72%. The model converged slower than all other models, and could not reach the same level of performance in 15 training epochs. The model would likely perform better, given more training time.

However, an interesting tendency is visible when analysing the Top-3 accuracy. The best video classifier, R(2+1)D, achieves a Top-3 accuracy of 97.15%, the best result out of all models. This finding hints at the fact that the temporal context of the video helps the robustness of the model: While single-frame classifiers make almost random predictions (the predicted class is not in the Top-3 predictions) in 9.5% of all cases, video classifiers only make such predictions in 2.85% of all cases.

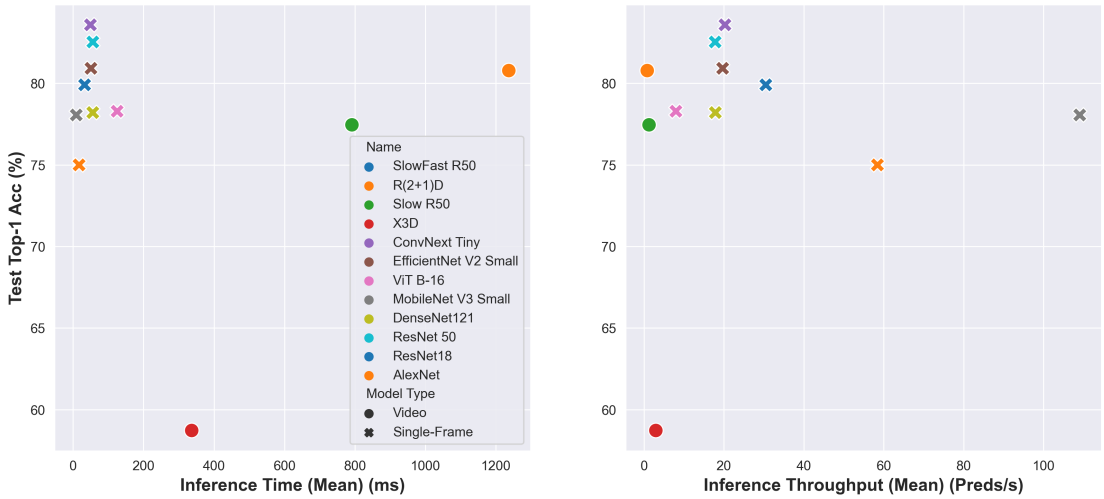


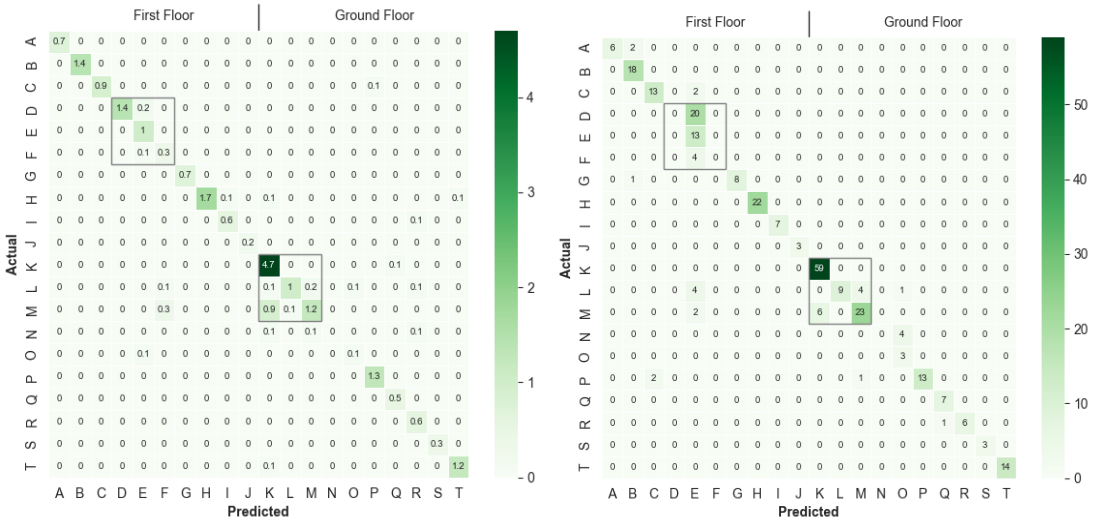
Figure 11: **Performance-Efficiency Trade-Off.** The Figure visualises the performance-efficiency trade-off for all models by plotting the relationship between the Top-1 Accuracy against the latency (inference time in milliseconds per prediction) and throughput (predictions per second). Each model is given unique colour, and the marker type indicates whether the model is a single-frame classifier (cross) or a video classifier (video).

5.2 Efficiency Analysis

In deep learning, more complex models generally outperform simpler ones if provided with sufficient training data and computing resources. However, there is a trade-off between model complexity and efficiency, especially when efficiency is critical, like when deployed on a mobile device.

The efficiency of the single-frame and video classifiers generally varies more than the performance - within and between the groups of singe-frame and video classifiers. The single-frame classifiers range from a throughput rate of 7.99 predictions per second (ViT-B-16) to 109.07 predictions per second (MobileNet V3 S). Three models (MobileNet V3 S, AlexNet, ResNet 18) can provide predictions in real-time (at least 30 predictions per second). The video classifiers require more computing and memory resources during inference and are less efficient. They are only able to provide predictions between 1.26 predictions per second (Slow R50 Family) and 3.01 predictions per second (R(2+1)D). However, as each prediction considers a history of frames, the predictions are more robust and valid for longer.

Given these findings, it is clear that performance gains can be achieved with an increased model complexity. However, small gains come at a high cost in terms of efficiency. Figure 11 visualises the performance-efficiency trade-off for all models by plotting the relationship between the Top-1 Accuracy against the latency (inference time in milliseconds per prediction) and the throughput (predictions per second). Depending on the specific deployment requirements, different models are more suitable. If low memory consumption and high throughput are critical, high efficiency models like MobileNet V3 S are most suitable. If the real-time inference is not critical but high accuracy is required, models like ResNet50 or ConvNext Tiny are best suited. If more robust but less frequent predictions are required, video classifiers like R(2+1)D can be valid options.



(a) ConvNext Tiny

(b) R(2+1)D

Figure 12: **Confusion Matrices.** The Figure shows the confusion matrix for (a) ConvNext Tiny and (b) R(2+1)D. The confusion matrix of ConvNext Tiny is normalised by a factor of 1/1000 for visual purposes. For both matrices, the entry at row i and column j shows the number of samples that belong to class i but were predicted to be class j. Gray rectangles indicate challenging subsets of classes, that are often confused. These are displayed in more detail in Figure 13 and discussed in detail in Section 5.3.2.

5.3 Understanding Model Behaviour

To better understand the complex dynamics, strengths and weaknesses of deep learning models in tackling the task of indoor classification, the two best performing models in each group, ConvNext Tiny and R(2+1)D, are analysed in more detail.

5.3.1 Confusion Patterns

Figure 12 shows the confusion matrices for (a) ConvNext Tiny and (b) R(2+1)D. Although the two models have very different architectures, they exhibit similar confusion patterns.

Both models show a tendency to confuse visually similar classes. The first example is the two corridors on the ground floor (Ground Floor Corridor 1 and Ground Floor Corridor 2, represented by letter L and M in Figure 12) and the atrium (Ground Floor Atrium, letter K in Figure 12). The corridor's architecture, interior design and lighting are very similar, making it more challenging for both models to distinguish between the two. Figure 13a shows the confusion between these classes for both models in detail. ConvNext Tiny tends to predict the atrium instead of the second corridor and regularly mixes up the two corridors. R(2+1)D is less prone to confusing the two corridors but also mispredicts the second corridor as the atrium in a few instances.

A second example is the three libraries (First Floor Library 1, First Floor Library 2 and First Floor Library 3, represented by letters D, E and F, respectively in Figure 12). The libraries are visually similar, as they share similar characteristics, like bookshelves, tables and chairs. Figure 13b shows the confusion between these classes for both models. ConvNext Tiny is more capable of differentiating the

three libraries but still confuses the first and second libraries. Weirdly, R(2+1)D cannot differentiate between the libraries and predicts the second library naively. This might be, because it is the most commonly occurring library in the training data, and the model is therefore biased towards this class.

The above failure cases highlight another intricacy of the models: While locations that are less present in the training data are generally handled well, the models tend to predict the majority class in cases of uncertainty.

5.3.2 Misprediction Cases

Exemplary mispredictions are analysed for both models confirm the above findings. Figure 14 displays six exemplary mispredictions for ConvNext Tiny (Figure 14a) and R(2+1)D (Figure 14b) each. The six mispredicted samples are grouped into three commonly found misprediction patterns:

Visual Similarities. This pattern describes mispredictions between visually similar classes. In the case of this study, there are several pairs or triples of classes (e.g. the two corridors, three libraries), which share many visual characteristics. The two samples in this group for ConvNext Tiny uncover another of such groupings: the coloured areas on the ground and first floor. The model is visibly less sure about the prediction, and the true class is typically the second most confident prediction. For R(2+1)D, the first sample shows confusion between the libraries, and the second sample shows confusion between the two corridors on the ground floor.

Location Transitions. This pattern describes mispredictions between classes that occur during the transition between areas. The first sample for ConvNext Tiny is a clip that just passes the door between two libraries. The model predicts the library before the door is passed, and the ground truth label is the library after the door is passed. The second sample shows a similar case, where the model predicts to be in the red area on the first floor, but the ground truth label is still the class before. This confusion pattern naturally arises because locations may not be sharply separable only from the imagery of a camera. For example, if the camera only captures the features of the location that lies a few meters ahead, it is likely that the model predicts the location that lies ahead, but the ground truth label is still the location that lies behind.

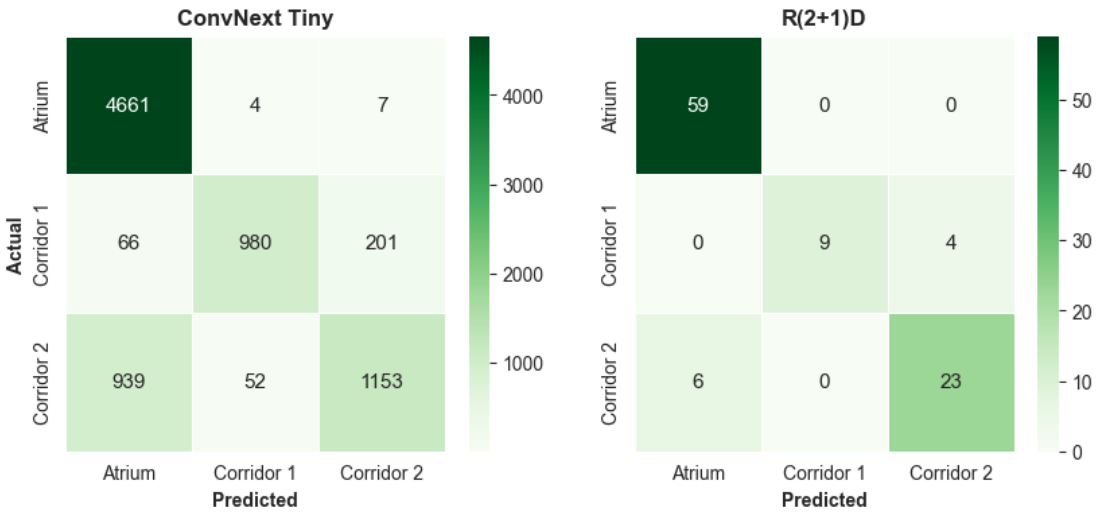
Environment Change. A third, interesting pattern was observed for both models in the area that was renovated in between the data collection periods for the training and testing split. Parts of Corridor 2 on the ground floor were renovated and repainted in blue. Both models fail to recognise this area and typically fall back to the two most common classes in the training data, which is the Atrium and Library 2.

5.3.3 Model Behaviour Analysis

Finally, to emulate the behaviour of the models in a real-world scenario, both ConvNext Tiny and R(2+1)D continuously predicted the location in all videos in the test split. As a result, the following additional qualitative observations were made:

Prediction Consistency. The single-frame classifier, ConvNext Tiny, was more inconsistent in its predictions than the video classifier, R(2+1)D. In complex cases, like the ones described in the previous section, ConvNext Tiny would often predict different classes in a short time span, leading to "flicker" in the predictions. This behaviour was almost absent in R(2+1)D due to the lower throughput rate and its temporal modelling capabilities, which allow the model to smooth out predictions over time.

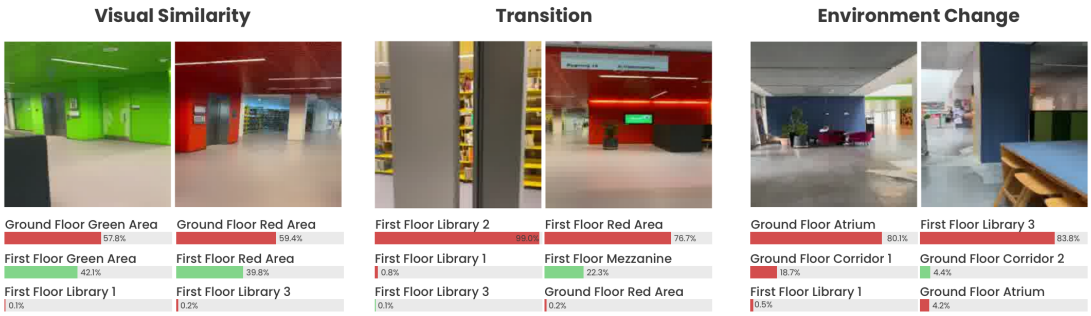
Training Data Bias. A bias toward the training data is present in both models. If features are not representative of a location but are overrepresented in the training data of that location, there is a chance for the model to learn these features as indicative of a class. This was the case for the



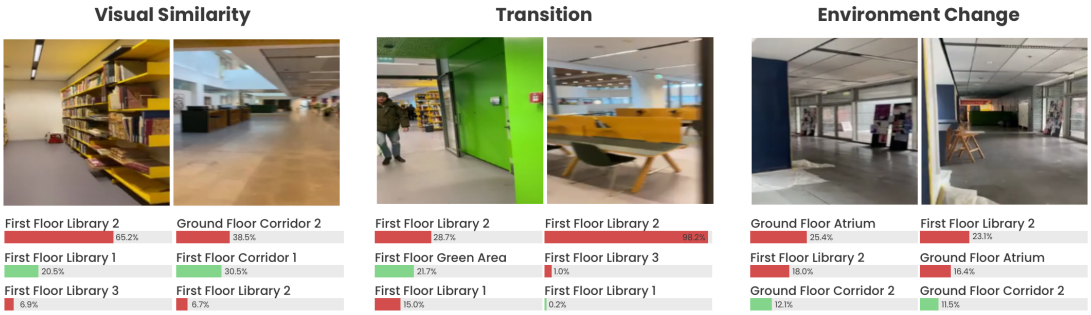
(a) Corridors

(b) Libraries

Figure 13: **Common Misprediction Patterns.** The Figure shows subsets of the confusion matrices that highlight classes that are commonly confused by ConvNext Tiny (left) and R(2+1)D (right). Figure (a) shows the confusion between the two corridors on the ground floor and the Atrium. Figure (b) shows the confusion between the three libraries on the first floor.



(a) Misprediction Patterns for ConvNext Tiny



(b) Misprediction Patterns R(2+1)D

Figure 14: **Mispredicted Samples.** Exemplary mispredicted samples for (a) ConvNext Tiny and (b) R(2+1)D. For each sample, the top 3 predicted classes are shown in order of confidence. The true class, if among the top 3 predictions, is highlighted through a green bar. The mispredicted samples are grouped in the three commonly observed misprediction patterns: (1) Visual Similarity (2) Transition and (3) Environment Change. For the clips from the video classifier, the first frame of the clip is shown.

libraries: Most video clips that were taken while walking through bookshelves were filmed in library 2, while the other libraries were filmed in a more open space. This led to models associating in-between bookshelves clips to library 2, even though they are also present in the other libraries.

5.4 Deployment on Mobile Devices

As a proof of concept, the most efficient and mobile-optimised model, MobileNet V3 Small, was deployed on a mobile device. The trained model was quantised to 8-bit float precision, converted to the `TorchScript` format, and finally deployed using the `PlayTorch` framework. The deployed model can be tested by downloading the PlayTorch app from the App Store or Google Play Store and scanning the QR code found in the README of this project's GitHub repository. After scanning the QR code within the PlayTorch app, the model will be downloaded and run locally on the device.

6 Conclusion

6.1 Limitations & Future Work

Arguably, the most significant limitation of this study is the small dataset used for training and evaluation.

The problem is two-fold: First, the relatively small training set leads to the models only seeing a subset of the possible angles, routes and natural variations in indoor spaces. As a result, consequences were visible in the misprediction patterns, as models often confused visually similar locations. Second, the small evaluation set might not represent the variation of visual inputs from an indoor location over a year. Therefore, this study's evaluation might not capture the potential weaknesses of the models in real-world deployment.

Therefore, a sensible next step would be to collect a larger dataset that captures the full variation of visual inputs from an indoor location over a long period of time. This would allow to investigate if more training data can overcome some of the misprediction patterns observed in this study. An interesting follow-up research question that speaks to the practicality of the proposed methodology is *how much* data is needed to achieve a certain level of accuracy; ideally, as little data as possible should be collected to reduce data collection and annotation cost while maintaining a high level of accuracy.

Finally, the detailed analysis of the mispredicted samples has revealed three frequent misprediction patterns: The fact that models depending on the modality of vision alone are unable to maintain high accuracy when drastic changes in the environment occur is a fundamental limitation of the proposed methodology, and likely cannot be overcome without major changes to the methodology. The second pattern, mispredictions at transitions between rooms, is also inherent to the task of room-level localisation, as it is sometimes hard to clearly distinguish when a transition occurs. Future work could investigate how to make transitions more smooth. Lastly, the confusion of visually similar classes is the problem that is most likely to be solvable. As mentioned, the first step would be to collect more data to allow the models to learn more discriminative features. Another starting point might be to directly incorporate the relative position of the rooms in the building, which could limit the number of possible predictions. For example, if a model is confident in predicting the current room, it could use this information to limit the number of possible predictions for the next frame if it has information about the current adjacent rooms.

6.2 Summary

This study has demonstrated that it is possible to train deep learning models and deploy them on mobile devices to provide reasonably accurate location estimates in indoor spaces.

Both single-frame classifiers that only use static frames as input and video classifiers that directly model the temporal information in videos, were capable of learning discriminative features from the visual input that allows them to localise a human agent in an indoor space from a set of rooms with a test accuracy of up to 83.59%. While the overall performance of models, also across groups, was similar, video classifiers were more consistent and robust in their predictions. However, this added robustness comes at the cost of a lower throughput rate, making the models less suitable for deployment on mobile devices.

A detailed analysis of the model's misprediction patterns and complex behaviours of unveiled three frequent failure modes: The models struggled with visually similar locations, transitions between locations and when significant visual changes, such as renovations, occur in the environment. Furthermore, both models were susceptible to biases in the training data, for example, if a feature that is not representative of a location is overrepresented in the training data of that location.

Overall, the results are promising and suggest that with a larger-scale and carefully designed dataset, it is possible to train models that can overcome the limitations observed in this study to provide even more accurate location estimates in indoor spaces. This could make the proposed method a viable alternative to existing indoor localisation systems, for example, in indoor navigation or augmented reality applications.

References

[1] N. Alsindi, B. Alavi, and K. Pahlavan. Measurement and Modeling of Ultrawideband TOA-Based Ranging in Indoor Multipath Environments. *Vehicular Technology, IEEE Transactions on*, 58:1046 – 1058, 04 2009.

[2] N. Alsindi and K. Pahlavan. Cooperative Localization Bounds for Indoor Ultra-Wideband Wireless Sensor Networks. *EURASIP Journal on Advances in Signal Processing*, 2008, 04 2008.

[3] U. Bandara, M. Hasegawa, M. Inoue, H. Morikawa, and T. Aoyama. Design and implementation of a Bluetooth signal strength based location sensing system. In *Proceedings. 2004 IEEE Radio and Wireless Conference (IEEE Cat. No.04TH8746)*, pages 319–322, 2004.

[4] J. Carreira and A. Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *CoRR*, abs/1705.07750, 2017.

[5] S. S. Chawathe. Beacon Placement for Indoor Localization using Bluetooth. In *2008 11th International IEEE Conference on Intelligent Transportation Systems*, pages 980–985, 2008.

[6] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. MonoSLAM: Real-Time Single Camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. *CoRR*, abs/1411.4389, 2014.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR*, abs/2010.11929, 2020.

[10] C. Feichtenhofer. X3D: Expanding Architectures for Efficient Video Recognition. *CoRR*, abs/2004.04730, 2020.

[11] C. Feichtenhofer, H. Fan, J. Malik, and K. He. SlowFast Networks for Video Recognition. *CoRR*, abs/1812.03982, 2018.

[12] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi. Escaping the Big Data Paradigm with Compact Transformers. *CoRR*, abs/2104.05704, 2021.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *CoRR*, abs/1512.03385, 2015.

[14] S. He and S.-H. G. Chan. Wi-Fi Fingerprint-Based Indoor Positioning: Recent Advances and Comparisons. *IEEE Communications Surveys and Tutorials*, 18(1):466–490, 2016.

[15] A. Howard, M. Sandler, G. Chu, L. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam. Searching for MobileNetV3. *CoRR*, abs/1905.02244, 2019.

[16] G. Huang, Z. Liu, and K. Q. Weinberger. Densely Connected Convolutional Networks. *CoRR*, abs/1608.06993, 2016.

[17] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *CoRR*, abs/1502.03167, 2015.

[18] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-Scale Video Classification with Convolutional Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[19] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The Kinetics Human Action Video Dataset. *CoRR*, abs/1705.06950, 2017.

[20] A. Khalajmehrabadi, N. Gatsis, and D. Akopian. Modern WLAN Fingerprinting Indoor Positioning Methods and Deployment Challenges. *CoRR*, abs/1610.05424, 2016.

[21] G. Klein and D. Murray. Parallel Tracking and Mapping for Small AR Workspaces. In *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 225–234, 2007.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *CoRR*, abs/2103.14030, 2021.

[25] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s. *CoRR*, abs/2201.03545, 2022.

[26] I. Loshchilov and F. Hutter. Fixing Weight Decay Regularization in Adam. *CoRR*, abs/1711.05101, 2017.

[27] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós. ORB-SLAM: a Versatile and Accurate Monocular SLAM System. *CoRR*, abs/1502.00956, 2015.

[28] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556, 2014.

[29] K. Sugiura and H. Matsutani. Particle Filter-based vs. Graph-based: SLAM Acceleration on Low-end FPGAs. *CoRR*, abs/2103.09523, 2021.

[30] A. Tahat, G. Kaddoum, S. Yousefi, S. Valaee, and F. Gagnon. A Look at the Recent Wireless Positioning Techniques With a Focus on Algorithms for Moving Receivers. *IEEE Access*, 4:6652 – 6680, 09 2016.

[31] M. Tan and Q. V. Le. EfficientNetV2: Smaller Models and Faster Training. *CoRR*, abs/2104.00298, 2021.

[32] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. *CoRR*, abs/2012.12877, 2020.

[33] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3D: Generic Features for Video Analysis. *CoRR*, abs/1412.0767, 2014.

[34] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. *CoRR*, abs/1711.11248, 2017.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. *CoRR*, abs/1706.03762, 2017.

7 Appendix

7.1 Reproducibility

All code and data used in this project is available on GitHub. The project's `README` file contains detailed instructions on how to reproduce the results of this project or just try out some of the trained models on an example video clip locally.

Further, the precise configuration, results and training logs of all experiments are available through the public Weights & Biases experiments.

7.2 Machine Specifications

Table 3 lists relevant hardware specifications of the remote server, that was used for training, evaluation and benchmarking of all models.

	Specification	Value
Sys.	Name	Linux
	Node	Desktop 24
CPU	Model	Intel(R) Xeon(R) CPU E5-2660 v3 @ 2.60GHz
	Architecture	x86_64
	Physical Cores	20
	Frequency	3.3 GHz
GPU	Model	NVIDIA GeForce GTX 1080 Ti
	Memory	11.2 GB
Mem.	Total Capacity	250 GB
	Avg. Used Capacity	~ 7.4 GB

Table 3: **Machine Specifications.** The Table shows relevant hardware specifications for the remote server that was used for conducting experiments within this study.