Multi-Person Radar Localisation Based on Vision Transformer

Alex Greenacre,w23044537

***Abstract*—Radar based human localisation is the ability to detect a humans location using object penetrating radars. This field has been shown to have a number of potential uses including a private cctv monitoring system, vital monitoring and to aid in search and rescue operations. This paper will explore the use of multi person human localisation through objects such as walls and doors in a search and rescue scenario proposing the use of a Vision Transformer (ViT) neural network alongside popular convolutional neural networks in this field to explore this models viability both in this scenario and for more general localisation uses in the wider field. Extensive experiments covering the effectiveness of the models were carried out exploring the models performance in various tests against a public dataset of over 60 different configurations, achieving distance loss figures of 8.3cm for the ViT model and 3.9cm for the ResNet model. The wider results from the experiments showed that a ViT model would not be viable in a search and rescue scenario against the current data size although one popular model in the field the ResNet model excelled in this situation showing it's viability in search and rescue and it's potential in the greater localisation field. Furthermore the results showed a need the field has for a large and varied public bench marking dataset for the field to excel as well as the ViT model to be further considered.**

***Index Terms*—Article submission, IEEE, IEEEtran, journal, LaTeX, paper, template, typesetting.**

I. Introduction

HAR as highlighted by [1] is the ability to recognise humans is one of the key areas of computer vision and machine learning before summarising that the two main questions of human interaction in this field is the recognition of the object and where it is. This could also be described as person identification and localisation and has became a focus point in the research community with potential uses of this technology leveraging the advantages of radar such as using the radars lack of identifiable information to implement a more private CCTV systems [2] or it's capability of surface penetration to detect humans through walls in situations such as search and rescue operations [3] [4] [5].

This paper will aim to explore the field of human localisation through radar and what further impact could be made to this field both in a wider context and when applied to search and rescue. This will be done by introducing a novel neural network not seen to radar localisation, the Vision Transformer model (ViT) that will be ran alongside other baseline models well known to radar based localisation and HAR as a whole in several experiments where the models will be compared against each other as well as the wider radar based localisation field using a stepped frequency continuous wave (SFCW) multiple input multiple output (MIMO) radar based dataset specifically designed for localisation in search and rescue environments. Contributions that the research will aim to contribute to are:

- The implementation of a ViT model for use in radar based localisation against a SFCW based radar dataset
- A exploration of ViT transformer when applied to search and rescue scenarios
- How a ViT model performs when compared previous to radar based localisation models

The results of these scenarios should help determine the models effectiveness when applied to search and rescue based scenarios as well as this the model should also be compared against the wider field of radar based localisation to further determine the viability of it use in this field in additional scenarios.

This paper was produced for the KF7029 Module for Northumbira University, Newcastle Upon Tyne .

Paper submitted August 28, 2024.

II. Literature Review

A. Person identification

A SFCW radar used by [6] to gather data on a participant in two separate positions (standing and siting) to detect a person when either in the open or behind a wall, this experiment produced a model that was able to detect a person with a high accuracy (98% in the open and 97% when behind a wall), although these are good results the experiment used a single person in a static position. [7] builds upon this taking data received from a SFCW radar and passing that data through a Resnet inspired convolutional neural network leading to a accuracy of over 96% when detecting if up to 2 people are present. Although this did give a lower accuracy than [6], [7] shows that a more advanced network could allow for better accuracy when detecting multiple humans and their posture.

Although some posture research focuses on identifying a pose as a classification labelling the posture in a set position [8], [9] explores the tracking of a humans skeleton using SFCW and a multi stage neural network consisting of CNN and GRU layers to generate a skeleton with a low error rate. This could allow dynamic posture positions to be monitored.

B. Person localisation

When investigating the localisation of people, [10] investigates using a IR-UWB radar to collect the data and find the location of people within a car using a multi layered perceptron model(MLP), this produces a model that can detect how many

people are in a car with accuracy up to 99.5%. although the model shows produced by [10] shows good results using classification this is not often used in localisation with some sort of co-ordinate based localisation used instead, this can be seen in [11] who uses errors such as mean error to show how far both models that they are comparing are from guessing multiple people's location based on their leg movement from LiDAR data. There has been research into localisation of a moving object with research by [12] who finds that they are able to use a series of uwb radars along with a cnn model to track a human as they walk between radar setups with a good accuracy, finding that the model worked at tracking a moving target with a mean square error of less than 30cm.

The tracking of multiple objects can be seen in by [13] where their research's into localisation found that they could use a FMCW MIMO radar to detect the location of up to 5 stationary people with a error of under 12cm this was later improved in a extended paper by [14] who uses a new pre-processing method called CHEAN to improve the performance of the CCN model used in the original paper gaining a lower error when finding the location of the participants.

C. Vital signs

More recently the aims of HAR through radar have started to expand past the localisation and recognition of humans and has started to branch out using the unique factors of radar to gather more information about the state of the person detected, monitoring factors such as their vitals in both breathing patterns or ecg. These methods of detection were explored by [15] who uses a IR-UWB and FMCW radars to explore the monitoring of vitals when different activates are carried out, this proves to be successful with [15] using a LSTM-CNN based model called A-FuseNet to good effect getting a accuracy of over 90% in all scenarios in the experiment and scoring higher than the models that [15] compares against. Although [15] does explore the vital signs and human recognition the paper does not explore the localising the position of the human.

These factors can be seen in research by [4] where they explore using a multiple input multiple output SFCW radar to collect localisation and vital data on single or multiple participants in multiple locations, positions and environments before trying to extract this data using a multi step algorithm to detect vital signs as well as each persons location in the room. Although [4] covered a large amounts of factors the results produced were very mixed, having found success when detecting both a participants breathing patterns and being able to localise a participants position with a 15cm error, Although the algorithm used had a high false detection rate when detecting the amount of people as well as the high error rate when the radar was used against a wall. Furthermore unlike [15], [4] only carried out research on stationary positions, although this is not an issue in the search and rescue scenario proposed it may mean that any model or algorithm trained on this dataset does not account for vital fluctuations and changes

D. Search and Rescue

The application of search and rescue in the radar field has been explored from a multiple of methods [3]proposes the use of using FMCW radar placed by drones at multiple points of a search and rescue site to detect a humans breathing patterns and using this to triangulate a persons position. [5] also explores the use of drones to identify respiratory signals except with the use of a SDR radar. both of these papers take a different approach to Schroth, *etal*. [4] Who instead of proposing air drone based detection instead chooses a track based drone, this gives a additional scenario to search and rescue based research opting for search and rescue operations in functional buildings rather than in rubble. This scenario also opens up the opportunity to allow for the use of research by the wider localisation field, as the drone will be driving up to walls in a similar fashion to current experiments [13] [16] [6]

III. Approach

A. Dataset

To carry out the experiments a publicly available database by [17] will be used, this database uses a SFCW MIMO radar to detect humans based on search and rescue scenarios. [17] is made up of over 60 scenarios covering the radar placement behind doors walls or in the open, the amount of people in the room with up to 5 people present and the co ordinates and posture of each person based on their distance from the sensor. As well as position and posture, vital signs for each participant were included with each scenario this includes the participants breathing patterns and ecg patterns.

All this information means that this dataset has potential for the goal of this paper, this is due to the large and varied scenarios allow for the proposed model to be trained an evaluated against a diverse dataset as a whole whilst also allowing a wide range of experiments where certain variables to be controlled (such as the posture or obstacle) whilst also retaining a large amount of data. As well as the strong potential of the dataset the academic literature surrounding the dataset itself shows that there is potential for the proposed model to further explore the performance of this models dataset and evaluate inconclusive areas such as the behind wall data and as highlighted in other search and rescue based research [4]

Although the dataset has potential there are drawbacks to using this dataset the first being that the wall used is self constructed, this as highlighted by [16] means that it may not be the best representation of an actual wall due to the lack of metallic objects such as wires or pipes. Furthermore the dataset does not account for different demographics of humans which could lead to bias in the model that is being created.

Also as shown in the datasets supporting research [4] the dataset is more focused on mathematical solutions to each scenario this means that although a number of varied data is present in the scenario, data distribution on the data may not be the best for training models with scenario such as ones with 4 people having significantly less data available than data with 1 or 5 people, furthermore some object types are completely missing from some categories such as scenarios with 5 people only being ran in a open environment, although the variation

in the data does mean that the models can be trained and compared this will only be against the full dataset and single person comparison may be distorted due the imbalance

1) Data structure: To use the data for training the data will be formatted so each feature will be split by the radars slow time sample, this represents a single full signal cycle with the data containing each frequencies return strength and time for all frequency steps as well as all antennas making up the mimo antenna. Formatting the data in this way will give the neural network models a training data of the following structure this will give the dataset a shape of (feature, frequency, antenna, time and strength of returned signal). This format gives a dataset of with a total of 118000 features.

B. Proposed Model – Vision Transformer

The new model being proposed is a vision transformer (ViT) model this model is inspired on the ViT model produced by [18], This model was first proposed to classify 2D images although with some modifications the model can be used to segment the radar signals response into segments, applying a token to the segmented data should allow the model to identify what segments contain humans and their location. Given the models success in radar based joint identification [19] as well as other classification fields [20] the ViT model could prove to have a large effect in the radar based field. Applying this model in the experiment will help determine how a effective a transformer model is when compared to the convolutional based models that are often used in HAR through radar.

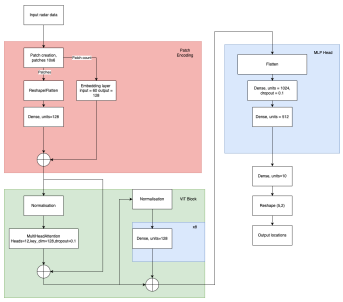

Fig. 1. Diagram of ViT model structure

To apply the model the following structure was implemented

1) Patch Encoding: The structure of the ViT model will follow that shown in Figure 1, the first part of structure will split the slow sample of the radar into 60 patches these will be split in a format of 10% per fast time sample and 20% of the antennas (10 fast time x 6 antenna), this will give each patch a size of 13,2,2. After each patch has been split the patch will be flattened to a 1D array and each patch will gain a position corresponding to where it would be in the original radar shape.

2) ViT block - Multi Attention Head: Once the patches have been encoded the encoded patches are then passed into a multi head encoder layer, this layer is based off the paper by [21] and focuses on checking multiple patches at once using the models heads, these heads are then concatenated into a vector before outputting, this when applied to the radar data it should allow for the model to view multiple patches of a radar slow sample and update the vectors of the embedded patches based off the what discoveries the attention head makes. This part of the block will maintain the 128 unit dimension set as well as this 12 heads will be implemented to reflect the lower bounds of the recommended heads as produced by [18] as well as this a dropout of 0.1 will be implemented to avoid reliance on certain connections

3) ViT block – MLP: The second part of the ViT block will focus on a sequence of mlp layers although originally proposed as one single hidden layer by [18] a sequence of 8 dense layers was implemented with the idea that it would help further identify the findings produced in the Multi Attention Head layer.

4) Flattened layer: As well as [18] the model will also implement a flattened layer before the mlp head as proposed by [22] this change will be made to help the MLP head fully capture all of the data and to better allow for localisation rather than classification.

5) Mlp Head and output: Once the data has been flattened it is sent through a mlp layer this layer consists of a two dense layer with units of 1024 and 512 that will take the tokenised data and send the updated data to a final output layers consisting of a 10 unit dense layer and a reshape that will resize the data into a distance and degree value for 5 people in a shape of (5,2).

C. Base Models

To compare results metrics will be used as a baseline to compare results against the novel ViT model

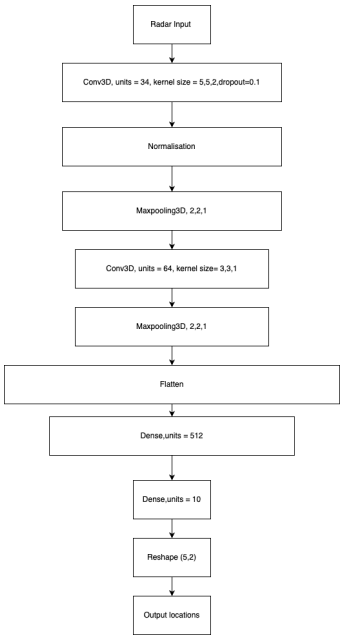

Fig. 2. Diagram of CNN model

1) Baseline model CNN: The CNN model is a AI commonly used in some form within the localisation field [13], [16] this model will be introduced as a base line to the novel neural network proposed allow the models performance to be measured against that that is used in the field, the model consists of two CNN layers the first with a unit size of 32

and a kernel size of (5,5,2)and the second with a unit size of 64 and a kernel size of (3,3,1) both layers are proceeded by pooling layers that half the data size each after layer, as well as this dropout and normalisation have been implemented between layers one and two to avoid over fitting or reliance on outliers.

2) Baseline model ResNet: Like the CNN model the ResNet architecture is a common model used within the field [7] and has proven to be a reliable model when applied in various forms within the wider HAR and object detection field [23] [24], because of this ResNet will be the second baseline model used to compare the novel model against. This model will use the in built keras_cv ResNet34 architecture based off the model created by [25].

D. Comparisons To The Wider Field

Alongside baseline scenarios comparisons to results produced by other researchers are also used, this will display the potential impact that these models may have against previous work. to do this two types of comparisons will be made Direct, against papers that have used the same dataset as this research and Indirect where papers have opted for a different radar or dataset.

1) Direct Comparison: Along with comparing to the wider field a comparison will also be done against research currently done against the dataset, this will be done by comparing against [4] a paper which used a MUSIC algorithm to determine the location of up to five people, this research shows a good performing model and as it has been carried out on the same dataset as our models should provide a direct comparison to the results of the experiments.

2) Indirect Comparisons: As the data set up is similar to the wider field comparisons can be made to other localisation papers that aren't necessarily search and rescue related to implement this the results of [13] and [7](model results 3 meters and under to match the furthest point in the dataset) will be used to compare the models against, this should give some indication of how the models perform against the wider field of radar based localisation.

E. Model Variants

As the models will all be trained using one model this means that all models will produce a fixed output of 5 people by 2 co-ordinates, this gives the opportunity of testing each models performance in two potential implementations

1) Standalone: This will be built to be implemented by itself (Figure 3) meaning that it will have to identify if 1 human is in the scenario or if 5 are and updates its output accordingly to illustrate the location of the humans present as well as clearly showing if humans are missing in the scenario

As this variant of the model allows for it to be ran in a standalone manner the model will be the default variant for all scenarios that do not require both variants of models.

2) As part of a network: Previous localisation research has lead to models that produce extremely accurate results of people present in the room with some models being as accurate as 99.7% [7], because of this a existing model could

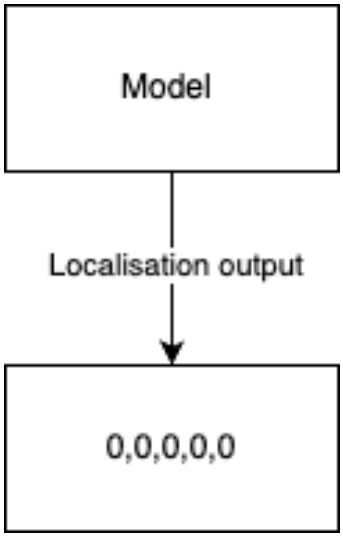

Fig. 3. Diagram of model workflow when ran by itself

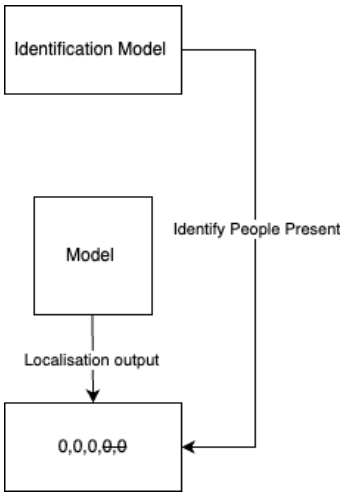

Fig. 4. Diagram of model workflow when ran with a identification model above the localisation model

be implemented to detect how many humans are present in the room and filter them out (Figure 4), this means that the output shape of the localisation model will be reduced after the model has ran depending on how many people the identification model filters out.

F. Loss functions and metrics

As the model will produce polar coordinates the error distance for the loss function will have to be converted to Cartesian this is due to the distance in polar measurements not being linear meaning that the loss in degrees further away from the radar will register the same as closer targets even though the distance is greater. To counteract this converting the loss formula should provide a linear loss and more readable loss in Cartesian(x and y co-ordinates) whilst also allowing the model to output polar co-ordinates which are more relevant to the scenario of search and rescue

G. Loss implementation

To account for the two separate variants of each model, the implementation of this loss function will occur in two separate formats, masked and unmasked.

1) Unmasked: To implement a standalone variant of the model the first model will be trained to assume that up to five people will be allowed in a room and that no person identifier model is also being used to detect, this means the identification will also be a part of the localisation model, to achieve this the model will consider the location 0,0 as an identifier that no person is present in the scenario

2) Masked: The second scenario will be trained assuming that a person identifier model will sit above it in the hierarchy and will indicate how many people will be present in the scene. To help simulate this a mask will be applied to the loss to filter out values produced by the model when the label has less than 5 people present.This means that the loss calculation will ignore predicted values if the corresponding validation labels are 0,0 and instead only focus on the loss of people identified in the scenario.

IV. Experiment Approach

To approach the evaluation of the ViT model the model,the base line models and the research highlighted will be ran in numerous scenarios to determine the models ability to detect humans both in a search and rescue scenario and in the wider field

A. Combined mean error

This scenario will evaluate the models performance overall using the combined loss of the ViT model when ran against all validation data, this will then be repeated for both CNN and ResNet baseline models to determine the ViT's effectiveness against models ran on the same dataset.As well as baseline models direct and indirect comparisons will also be made to wider literature to compare the models to localisation in other areas of the field. By carrying out this experiment it should give a clear metric of how well the model localises a humans position and how well it does this when compared to other attempts or models.

B. Set person count

To further explore the models effectiveness the models effectiveness will be explored against each amount of people present in a scenarios, organically this scenario would compare against [13] to determine if the model follows the same trend of a slowing decreasing accuracy as more people are added to the scenario but due to data distribution it would unfeasible to give this comparison. Instead the model will be compared against the baseline models to offer a better understanding of what models it excels in and if the other models follow the same trend. furthermore comparison against [4] can be made as a direct comparison to the same dataset.

C. Single Deployment vs Deployment With Identification Model

To explore the effect that identification as well as localisation may have on a models performance this scenario will compare two variants of the same baseline models against each other, the mask(part of a network) variant and the mask-less (standalone) variant to determine if a model can be deployed to both identify and detect people in a scenario or if it's more effective in a network. To carry this out the entire validation dataset will be used on both variants of each model to highlight the performance difference between both types of models as well as baseline models.

D. Object Performance

This scenario will consider the impact of the models performance against material blocking the radars view, this is due to multiple research finding that their models often performed worse against thicker materials [13] [4], due to this it would be beneficial to compare the novel model proposed against the different materials in the dataset (Open, Door, Wall) to verify what impact these materials will have on our model performance. To carry this out scenarios containing two and three people will be used, this is due to both scenario types having the same amount of object based scenarios and variety guaranteeing a even data distribution. this data will then be used to compare all baseline models to evaluate if they follow the same trend as [13] [4]

E. Model Training

The models will be trained against the full dataset in a train/test split of 70/30 will be used and will all be trained on 150 epochs once these models have been trained the model will be ran against all scenarios using the same trained models in each experiment. To carryout the training a Nividia Geforce GTX 3060 Ti 8Gb will be used.

V. Results

The results produced from the experiment cover the findings from over 30 separate tests covering 4 main scenarios

A. Scenario 1 – Combined localisation

Model	Distance error (cm)
CNN	31.1
ResNet	3.9
ViT	8.3
MUSIC Algorithm [4]	15
CEHANet [13]	3.8
MSCAM-ResNet [7]	9-13

TABLE I
TABLE CONTAINING THE COMBINED AVERAGE FOR THE VIT AND BASELINE MODELS AS WELL AS RESULTS FROM THE GREATER RESEARCH COMMUNITY

The results of the scenario (Table I) shows that Resnet and ViT outperform the research into the search and rescue by [4] with both models outperforming the MUSIC algorithm by 40% or more. Although the combination SCFW and ViT/ResNet

can be seen to outperform work in the wider field producing more accurate results when compared to the MSCAN-ResNet implementation by [7] although only ResNet can be compared to CEHANET [13] producing only marginally worse results.

B. Scenario 2 – Localisation per person

Fig. 5. Line chart displaying model results per person present in scenario

The results displayed in Figure 5 indicate, the ViT & CNN models are consistently outperformed by the resnet model for every scenario, Although this is the case ViT still outperforms the direct comparison of the MUSIC model by [4] in every scenario except for 3 people in which it is closely matched. Furthermore the ViT and ResNet models do not seem to follow the trend set by [4] giving spikes in distance error when moving to scenarios 2 & 3 which contain more object related data and not drastically increasing on scenario 5 which contains no object blocking.

Although the results of the tests can be compared to research directly linked to the paper comparisons cant be made to the wider field of radar. This is because the results are clearly effected by the different variations that are present in some scenarios compared to others, this can be seen in the comparison between scenarios containing two people and scenarios with five, as two containing scenarios covering different objects types such as radar behind doors and walls where scenarios with five contain no blocking object scenarios.

As well as comparing against existing research data distribution means that no conclusion can be clearly determined for the impact that increasing the person count could have on the models.

C. Scenario 3 – Masked values vs Unmasked

	Mask	Mask-less
CNN	65.2	31.1
ResNet	21.7	3.9
ViT	21.8	8.3

TABLE II
MASK AND MASK-LESS COMPARISON(CM) FOR EACH MODEL

As shown by Table II the results of this experiment showed very little difference between res net and ViT model with both model scoring similar results, however Resnet shows a much higher accuracy when ran with no mask, these results expand on what was previously identified showing that although Resnet may be better performing when needing to also identify the model both models are equal when localisation is the only goal of the model.

Model	MSE	MAE
CNN	31.1	51.8
ResNet	3.9	5.9
ViT	8.3	17.1

TABLE III
TABLE SHOWING MSE AND MAE FOR EACH BASELINE MODEL

Although each models performance could still be measured the results of this experiment did not match the expected results, this is due to the model with a unmasked loss scoring almost 6 times lower than that of a masked model. To explore this variation in difference, two factors could be applied to explain the results. The first being the hypothesis that the models require information about missing humans from the scenario to further identify what is and isn't human, as unmasked scenarios are exposed to humans not being present in its loss function it could learn to better identify and localise based on this additional dimension. The second hypothesis based around the accuracy of the models as the maskless model has gotten so good at determining the identification of humans in or out of the scenario the low errors surrounding this have started to effect the loss, this could be seen in the difference in mse to mae. As mse is focuses on high error scenarios squaring loss under one lessens its impact this leads to a scenario where the more accurate values produced the lower mse will be when compared to more linear loss such as mae. This can be seen in the results for scenario 3 although this does occur on both model types meaning that this may not be the main influence

1) Scenario 3.5 - Hypothesis exploration: To explore the results in greater depth every scenario was compared against by group number in the same way as scenario two except each scenario was ran using the masked and unmasked, loss these were then compared against to find how the implementation or subtraction of zeros based on the person count would impact loss difference between the two variations of models. As the main focus of this scenario is to view difference the data distribution issue will not effect the results in the same way as it did when comparing model effectiveness per person.

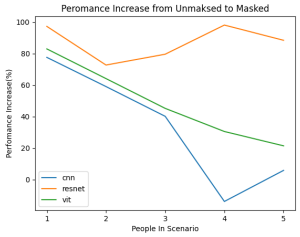

Fig. 6. Line chart showing each models performance against it's maskless and masked variant as people present in scenarios increases

The results of this show the link between CNN and ViT models to identification(Figure 6). As the amount of people in the scenarios increase the performance increase form both models can be seen to decrease, this helps support the hypothesis that localisation accuracy could be effected by the models ability to identify humans in mask-less scenarios, although this is the case for both CNN and ViT model, ResNet does not

follow this trend showing no correlation between the need to identify less humans as not present (0,0) and the accuracy of the model.

D. Scenario 4 – Impact Of Objects

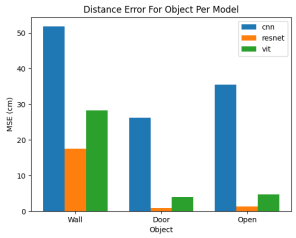

Fig. 7. Bar of model results against each object type

With the experiment concluded the results shows a clear trend in all three models when the introduction of a wall was used in the scenarios, with all models doubling their distance errors once a wall was introduced, this results correlates to the findings of other research done against this dataset with [4] finding detection of object through walls in these scenarios to be impossible with no scenario being under a error of 30cm Furthermore the wider field of radar also experience this issue with FMCW [13] also experiencing issues when applied to thicker walls validating that this issue may not just be isolated to SCFW based radars. Although the models shows inefficacy at dense objects such as walls there is success in objects with smaller density, this can be seen in the results against an open area and doors which shows that very little changes against the models performance with all models slightly outperforming themselves in door scenarios vs their performance in free areas.

Vit model performance in this scenario is shown to be slightly worse than that of Resnet although the model still outperforms the direct comparison in [4]producing lower results than the music algorithm

VI. Discussion

A. Data distribution

Given the poor distribution of the dataset it has not been possible to carry out scenarios in a preferred manner this can be best seen in the data collected in the individual person count in each scenario, this meant that no comparison could be made to the wider field limiting the ability to look for certain trends within the model. As well as this the exploration of an objects effects had to be cut down to two scenarios as they had the best distribution for this scenario. As well as distribution only five participants where used within this dataset and in turn this experiment, although this could be expected due to the early nature of this field the models will not exhibit a complete understanding of human demographics, this could lead to a bias where localisation can only determine locations of certain body types, this in a search and rescue scenario could be catastrophic if the user is neither identified or located.

1) Public Datasets: As highlighted in [26] [16] there are no well known or used public datasets that have been produced to accommodate radar based localisation with neural networks as it's focus, this means that a large amount of different papers found produced their own differing private datasets for their model

2) Dataset creation approach: Following other research in the field and creating a custom dataset would be a viable approach in elevating the problems presented by the data distribution as considerations such as data distribution, data variation and participant demographics could be considered. Although carrying out this data collection for the project could be possible with [16] [13] [7] [10] all carrying out their own data collection to build datasets more suited for neural networking learning, a lack of skill and resources in the radar field meant that the time it would take to learn, set up and enact a robust radar data collection the time frame supplied for this project was not viable.

B. Model performance – indirect comparisons

Another effect of data collection also being a part of the research in the field is that there is no set standard for how data could be displayed, for example [13] uses one scenario for validation before moving onto a completely separate one in a different room for additional validation. This differentiates widely from the approach used in [16] making cross comparison in papers harder to determine, although this is still possible to compare. Implementing a more standardised dataset in the future would mean that this comparison could be more accurate

C. Model structure

To avoid the impact of the localisation issues a model structure could have been used that trained and produced models only focused on one specific number for the amount of people present this would bypass the issues with the 0 labels that have been present in this research and instead allow the model to produce results focused only on one type of output potentially increasing the models performance.

D. Localisation bias

Given the results of scenario 3 point 5 that it is clear that CNN and ViT are linked to the labelling of missing humans from a scenario as their performance against the masked scenario with no identification present decreased as their was less missing participants in each scenario, this means that there is a clear link between the models performance against when ran by itself compared to when it is used with a person identifier filtering the results. This means that the identification of these missing participants is directly effecting the models accuracy against the position of people in the scene

Although this calls into question the validity of maskless variations of these models the high accuracy against the identification that a human is not present in this scenario. This further reinforces the results found by [7] and [10] that radars are highly effective at detecting and counting the amount of humans in a scenario

Furthermore Resnet does not follow this trend, this means that it its low accuracy in mask scenarios is not influenced by its ability to distinguish humans present in the scenario meaning that maskless scenarios covering both combined and individual results can still be considered using this model. Although not proven one reason that ResNet still outperforms using a maskless loss could be explained by the additional factor of the human not being present providing means for greater accuracy

E. ViT

Given the ties to identification, the results of the ViT model when ran as a single model (without a mask) cannot be confidently stated without the acknowledgement that its accuracy used to localise humans is influenced by ability to detect the amount of people in the area. Because of this the models effectiveness when ran in maskless such as scenarios one and two cannot be considered as concrete results. This means that comparison's against search and rescue scenarios and localisation can only be done with the masked results which show the ViT model as being worse than the results found in both the direct dataset and in the wider field.

Although there are issued with the implementation of the ViT model when unmasked the masked version does perform exactly the same as ResNet indicating that the model is as effective as currently used models in the field in some scenarios. Additionally ViT is known to be more effective the more it is scaled up as highlighted in both [26] and [20] meaning that the model could outperform current convolutional based models such as the basline ResNet model when applied to a larger dataset

F. Resnet performance

Although ViT performance has lacked ResNet has shown itself to be a strong option when considering models to use in localisation this can be seen in all results gathered as it out performed or equalled the ViT model in every scenario. Furthermore its rejection of the Identification trend shows that a standalone version of resnet is not only possible but it can detect humans to a high accuracy, outperforming the results found by the MUSIC algorithm by 60%+ percent [4] as well as producing similar results to the wider field such as the metrics found by [13] [7]. Giving these results it is clear that the ResNet model be chosen as the preferred model when carrying out search and rescue based human localisation using the SCFW and has shown that this model combination not only to be a viable option when considering this application but also in the localisation field as a whole

G. Object Impact

Although the dataset for exploring the impact of object density on the radar had to be cut down the experiment till produced some significant results, the main one being that thickness of materials impacted the performance of the models with a significant increase of over 100% on all models when switching from a door to a thicker concrete wall, this trend reinforces the findings of other work [13] [4] that as the thickness increases the performance of the radar decreases. Aswell as following the trends of the wider field the experiments did produce results that stated a slight increase performance when blocked by a door, this could show that this kind of material has no effect on the functioning of the radar. Overall this experiment supports the outcome suggested in [4] that for the models to be effective in a search and rescue scenario the radar should be placed against a low density object such as a door or in a open space. Although this paper was able to explore the impact of object density on the performance of the radar it should also be noted that trends surround the change of object materials in the wall may cause further performance impacts such as that which was found by [7]

VII. Conclusion

Based on the data collected ViT model cannot currently be recommended as an accurate alternative to current implementations due to its poor performance when compared to other research both in the search and rescue field and in the wider radar localisation due its reliance on person identification obscuring any results and when this this factor has been removed its has displayed poor performance. Although it cant be recommended in currently, the close link to resnet performance in pure localisation scenarios such as that of scenario 3 as well as it's ability to perform in high data scenarios means that the model could be considered in the future.

Although findings show that the ViT models performance currently lacks the results show that the ResNet model can localise humans in a search and rescue scenario with high accuracy whilst rejecting identification influence on accuracy. Furthermore ResNet not only preforms better than the direct research done on the dataset but it also outperforms wider research

For the field to move forward there must be a larger more diverse public dataset created for SCFW based radars for the specifically for the implementation of neural networks and SCFW to be fully utilised in localisation scenarios both as a whole and in more specified scenarios such as search and rescue. The existence of this dataset would remove a number of limitations that have been implemented on this project and would allow for models that could be built without compromising structure and scenarios due to data issues. Furthermore the creation of this dataset could help the comparison of model and their performance, this has been seen in other fields such as the BraTs challenge datasets [27] where researchers can use a set dataset to further push the capabilities of neural networks in that specific field [28]

As well as search and rescue the models could be applied to additional areas, a prime example of this could be a private cctv based monitoring system, this has been explored by [2] finding this scenario to be possible. The findings of this report shows some viability of this technology when SCFW and neural networks are applied, showing potential of this use in more static situations such as busses, car, metros or lifts. Should technology explored in radar based in vital recognition

such as that proposed by [15], be applied these technologies could be further expanded upon to monitor peoples health in these scenarios.

This research has shown the potential of ViT models in radar location and given a larger more diverse dataset future work on radars using ViT models should be carried out to produce more accurate models.

REFERENCES

[1] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," *Frontiers in Robotics and AI*, vol. 2, 11 2015.

[2] W. Butler, P. Poitevin, and J. Bjomholt, "Benefits of wide area intrusion detection systems using fmcw radar," in *2007 41st Annual IEEE International Carnahan Conference on Security Technology*, pp. 176–182, 2007.

[3] C. S. Chaves, R. H. Geschke, M. Shargorodskyy, R. Brauns, R. Herschel, and C. Krebs, "Polarimetric uav-deployed fmcw radar for buried people detection in rescue scenarios," in *2021 18th European Radar Conference (EuRAD)*, pp. 5–8, 2022.

[4] C. A. Schroth, C. Eckrich, I. Kakouche, S. Fabian, O. von Stryk, A. M. Zoubir, and M. Muma, "Emergency response person localization and vital sign estimation using a semi-autonomous robot mounted sfcw radar," 5 2023.

[5] B. P. A. Rohman, D. Kurniawan, C. B. Ali Wael, and A. Subekti, "Toward a compact and reconfigurable radar-uav for remote vital sign detection," in *2022 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, pp. 68–72, 2022.

[6] A. Kılıç, İsmail Babaoğlu, A. Babalık, and A. Arslan, "Through-wall radar classification of human posture using convolutional neural networks," *International Journal of Antennas and Propagation*, vol. 2019, pp. 1–10, 3 2019.

[7] J. Lin, J. Hu, Z. Xie, Y. Zhang, G. Huang, and Z. Chen, "A multitask network for people counting, motion recognition, and localization using through-wall radar," *Sensors*, vol. 23, p. 8147, 9 2023.

[8] H. Cui and N. Dahnoun, "Human posture capturing with millimetre wave radars," in *2020 9th Mediterranean Conference on Embedded Computing (MECO)*, pp. 1–4, 2020.

[9] Y. Xiang, Q. Tang, G. Cui, S. Guo, Y. Jia, and C. Chen, "Visualization of human posture based on radar time-frequency spectrogram," in *2021 CIE International Conference on Radar (Radar)*, pp. 3074–3077, IEEE, 12 2021.

[10] S. Lim, J. Jung, S.-C. Kim, and S. Lee, "Deep neural network-based in-vehicle people localization using ultra-wideband radar," *IEEE Access*, vol. 8, pp. 96606–96612, 2020.

[11] M. Guerrero-Higueras, C. Álvarez Aparicio, M. C. Calvo Olivera, F. J. Rodríguez-Lera, C. Fernández-Llamas, F. M. Rico, and V. Matellán, "Tracking people in a mobile robot from 2d lidar scans using full convolutional neural networks for security in cluttered environments," *Frontiers in Neurorobotics*, vol. 12, 2019.

[12] C. Li, E. Tanghe, J. Fontaine, L. Martens, J. Romme, G. Singh, E. D. Poorter, and W. Joseph, "Multistatic uwb radar-based passive human tracking using cots devices," *IEEE Antennas and Wireless Propagation Letters*, vol. 21, pp. 695–699, 4 2022.

[13] C. Wang, D. Zhu, L. Sun, C. Han, and J. Guo, "Real-time through-wall multihuman localization and behavior recognition based on mimo radar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[14] D. Zhu, C. Wang, C. Han, J. Guo, and L. Sun, "Twlbr: Multi-human through-wall localization and behavior recognition based on mimo radar," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, pp. 3186–3191, IEEE, 12 2022.

[15] X. Yang, X. Zhang, Y. Ding, and L. Zhang, "Indoor activity and vital sign monitoring for moving people with multiple radar data fusion," *Remote Sensing*, vol. 13, p. 3791, 9 2021.

[16] W. Wang, N. Du, Y. Guo, C. Sun, J. Liu, R. Song, and X. Ye, "Human detection in realistic through-the-wall environments using raw radar adc data and parametric neural networks," 3 2024.

[17] C. A. Schroth, C. Eckrich, S. Fabian, O. von Stryk, A. M. Zoubir, and M. Muma, "Multi-person localization and vital sign estimation radar dataset," 2023.

[18] 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," 10 2020.

[19] Z. Zheng, D. Zhang, X. Liang, X. Liu, and G. Fang, "Radarformer: End-to-end human perception with through-wall radar and transformers," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2023.

[20] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Applied Sciences*, vol. 13, no. 9, 2023.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 6 2017.

[22] K. Salama, "Image classification with vision transformer," 1 2021. Accessed: 24/08/2024. Avalible: https://keras.io/examples/vision/image_classification_with_vision_transformer.

[23] P. Yan and W. Shao, "A comparative study on repvgg and resnet for monocular 3d object detection," in *2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–6, 2022.

[24] Z. Sharifisoraki, M. Amini, and S. Rajan, "Comparative analysis of mmwave radar-based object detection in autonomous vehicles," in *2024 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1–6, 2024.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[26] G. Zhang, D. Zhang, Y. He, J. Chen, F. Zhou, and Y. Chen, "Multi-person passive wifi indoor localization with intelligent reflecting surface," *IEEE Transactions on Wireless Communications*, vol. 22, pp. 6534–6546, 10 2023.

[27] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[28] M. Ghaffari, A. Sowmya, and R. Oliver, "Automated brain tumor segmentation using multimodal brain scans: A survey based on models submitted to the brats 2012–2018 challenges," *IEEE Reviews in Biomedical Engineering*, vol. 13, pp. 156–168, 2020.