Transformer Network for Grouped Target Counting Tracking with a 24 GHz MIMO FMCW Radar

Wang, D.; Yarovoy, A.; Fioranelli, F.

DOI
10.1109/RadarConf2559087.2025.11204984

Publication date
2025

Document Version
Final published version

Published in
Proceedings of the 2025 IEEE Radar Conference, RadarConf 2025

Transformer Network for Grouped Target Counting & Tracking with a 24GHz MIMO FMCW Radar

Dingyang Wang, Alexander Yarovoy, Francesco Fioranelli
Microwave Sensing Signals & Systems (MS3) Group, Department of Microelectronics, TU Delft, The Netherlands
{D.Wang-6, A.Yarovoy, F.Fioranelli }@tudelft.nl

2025 IEEE Radar Conference (RadarConf25) | 979-8-3315-4433-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/RADARCONF2559087.2025.11204984

***Abstract*—The problem of radar-based multi-target tracking for indoor human monitoring is considered. Tracking and counting the number of people moving as a group is particularly challenging as multiple individuals are close together and their radar signatures are mixed. A transformer-based classification approach for counting the number of grouped people is proposed. The Neural Network model is trained with selected features from the spatial domain and Doppler frequency domain, which are concatenated over multiple frames to form a sequence for the transformer network. Compared to statistical classifiers, the self-attention mechanism allows transformers to capture feature long-term dependencies. The proposed classifier is integrated into a tracking pipeline in order to monitor the position and number of people in the grouped targets. The method proposed is experimentally verified using a 24GHz Frequency Modulated Continuous Wave (FMCW) radar with 250MHz bandwidth. Despite the relatively coarse range resolution, the proposed method achieves 92.5% accuracy in these initial tests. Furthermore, the method performances and related accuracy is analyzed according to various parameters.**

***Index Terms*—Multi-target tracking, grouped target, classification, transformer network.**

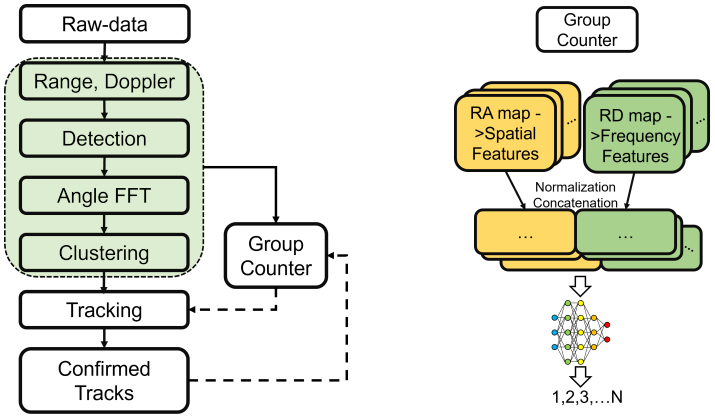
I. Introduction

Human tracking [1]–[5] is an active research area which had started from vision based sensors. More recently, radar sensors have been employed for the same task. Compared to vision based sensors, radars do not have severe privacy issues and are independent on ambient light conditions, which can offer tracking and monitoring capabilities continuously. Additionally, multi domain measurements can be obtained from a radar sensor, such as range information, radial velocity information via the Doppler effect, and intensity of the echo from different body parts. Besides this, the angular information is also available with the use of multiple input multiple output (MIMO) antenna array technologies in the most recent radar sensors. All these capabilities, combined with relatively low-price of radar, are promoting their usage for indoor human tracking and counting the number of people present in an environment.

In the context of human monitoring, a 'grouped target' can be considered as a cluster of people sharing neighboring locations and moving together [4]. The grouped target behavior is common in daily life, considering for instance a couple of friends moving to the elevator or a family walking together to cross the road, just to give two examples. From the radar point of view, it is rather challenging to distinguish each person within the cluster. Usually, there is not enough angular resolution due to the limited number of MIMO channels in relatively simple and cheap radars used for these use cases, and the cross-range resolution degrades at positions further away from the boresight. As a result of this challenge, in many radar research found in the literature, the participants being tracked tend not to come too close to each other, or a more general descriptor of 'crowdness' is given rather than a precise estimation of the number of people. Nevertheless, some features have been proposed to help classify the number of people clustered into a grouped target. These include for instance those extracted from range-azimuth maps [1], cadence velocity diagrams (CVD) as periodic representations of micro-Doppler signatures [4], or wavelet decomposition as an alternative [6]. The use of wavelets may present advantages compared to more conventional Fourier analysis of the periodicity of micro-Doppler signatures, improving resolution and using more localized frequency bands.

After features are extracted, there are different methods to predict the number of people in a group. Most of the previous research considers this as a classification problem, as the target variable to be estimated is inherently discrete. Otherwise, additional actions need to be taken such as rounding float value to integer if considering this as a regression problem. In our previous research [6], the Support Vector Machine (SVM) [7] classifier was used to solve this problem, and was integrated into a processing pipeline to perform simultaneously tracking of the different groups of people alongside counting their members.

However, SVM classifiers require a fixed length of inputs and are unable to model connections and dependencies of the feature values over multiple frames. This is a problem in practical scenarios where missed detection of people could happen during the tracking phase. The subsequent missed data and features would then need to be handled manually, such as with manual cropping or via forms of interpolation. For this reason, in this paper a transformer-based network is proposed and designed for the problem of classification of the number of people in grouped targets. Compared to SVM, transformer model is flexible in dealing with variable lengths inputs and can inherently manage long dependencies in the data and multi-class problems. The classification step to count the number of people is performed together with their tracking in a unified processing pipeline. Features based on the wavelet decomposition method in [6] and features from range-azimuth maps are used. The performance metric used is the OSPA



(a) Tracking pipeline (b) Grouped target counter

Fig. 1. Processing pipeline for tracking, integrated with grouped target counter

(Optimal SubPattern Assignment) [8], which is suitable for the assessment of multi target scenario. Also, OSPA includes the position error component, and a cardinality error component to quantify errors in estimating the number of people present.

The rest of the paper is organized as follows. In Section II, the feature extraction is presented. Then, the proposed transformer-based network for classification with times-series sequence input is presented. The experimental setup is shown in Section III, the results and performance analysis are provided. Finally, the paper is concluded in Section IV.

II. Proposed Method

In this section, the pipeline used for tracking and counting (Fig. 1) is described and depicted. In general, the pipeline includes a transformer based group target counter in addition to a more conventional tracking pipeline [9], [10]. On the left-hand side, there is a tracking block diagram which includes the steps from detection in the range-Doppler maps to clustering and tracking itself. After the classification for the grouped target, the number of people will be passed to the tracker to obtain the metric. The right-hand side shows the group counter with steps from feature extraction to classification, where each class corresponds to a number of people present in the group.

It is important to notice that the two blocks operate together into a unified pipeline. Essentially, the group counter block takes current features from the range-azimuth domain and Doppler frequency domain of data belonging to each specific track confirmed by the tracking block. As the track is maintained over time, a history of feature values for that specific tracked target is created for the classifier. Conversely, the classifier from the group counter block provides input to the tracking block in terms of the number of people present.

A. Feature extraction

In this work, a manual feature extraction approach is used. Compared to passing micro-Doppler signatures as images to a CNN network for pattern extraction, the manual method is easier to interpret and explainable. Spatial features derived from range-azimuth maps are extracted as expected to be related to different numbers of people observed. The spatial features are extracted by averaging within a window time. Specifically, the spatial features from range-azimuth maps are the width occupied in azimuth axis, length occupied in range axis, mean value of angle bins, median value of angle bins, variance of angle bins, number of different angle bins occupied, mean value of angle profile, median value of angle profile, variance of angle profile and number of pixels occupied in RA map.

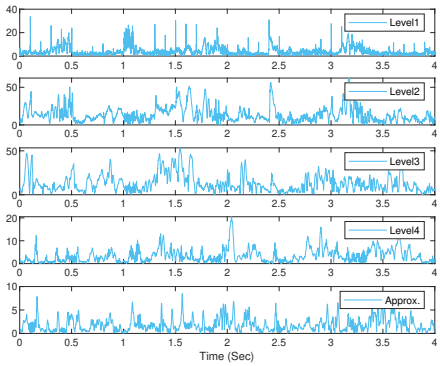


Fig. 2. Example of MODWT decomposition signals from the data for 1 target walking straight.

In order to obtain frequency domain features related to the velocity components of the people in the field of view, a sliding observation window is applied. Specifically, frequency features are extracted from different levels of wavelet decomposition as shown in Fig. 2. The Maximal overlap discrete wavelet transform (MODWT) [11], [12] is used for this purpose. First, the time series data to be analyzed is denoted as $X = \{x_t\}$ $t = 1, ...T$. The jth level wavelet and scaling filter are denoted as $\{\tilde{h}_{j,l}\}$ and $\{\tilde{g}_{j,l}\}$ respectively. The scaling and wavelet coefficient can be computed as follow:

$$\tilde{V}_{j,t} = \sum_{l=0}^{T-1} \tilde{g}_{j,l} x_{t-l \bmod T} \quad (1)$$

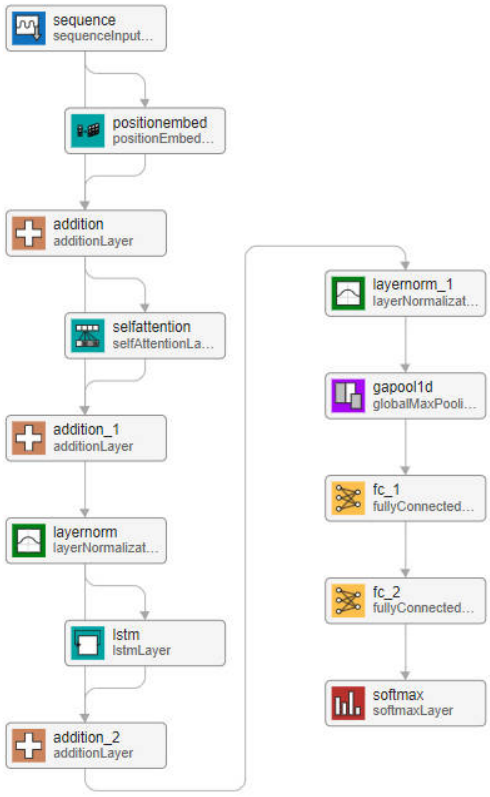


Fig. 3. Proposed transformer-based neural network architecture designed and implemented in MATLAB.

$$\tilde{W}_{j,t} = \sum_{l=0}^{T-1} \tilde{h}_{j,l} x_{t-l \bmod T} \tag{2}$$

while $j = 1, 2, ..., J$ is the level of wavelet decomposition, here $J = 4$. Compared to the conventional wavelet method, MODWT avoids excessive down sampling and prevents the data size to be too small when increasing the number of levels. Also, the frequency information is aligned with time. Overall, the Doppler frequency domain features for each level include the variance, standard deviation, mean value, median value, root mean square value, skewness value, kurtosis value, and entropy value.

B. Network for classification

The transformer network was originally introduced in [13]. The network architecture relies on self-attention mechanisms to model dependencies in sequential data. Recently, this kind of networks is adopted for classification tasks across multiple domains such as vision [14], and time series data [15]. Inspired by previous research, we designed a transformer-based network for people counting, as illustrated in the architecture in Fig. 3. The LSTM layer is primarily focused on capturing sequential dependencies, whereas the transformer architecture offers improved parallelization and pattern extraction capabilities [16].

TABLE I
JOBY (FORMER INRAS) FMCW RADAR PARAMETERS

FMCW radar model	RadarBook2 (RBK2)
Operating frequency	24 GHz
Sweep bandwidth	250 MHz
ADC sampling rate	120 ksps
ADC samples	56
Up chirp duration	467 μs
Chirp repetition interval	483 μs
Number of chirps in a frame	90
Slow-time sampling frequency	10 Hz
Number of TX & RX channels	2 x 8
Antenna horizontal 3 dB beamwidth	76.5°

The proposed network comprises 13 layers in total and is fed with features extracted from range-azimuth maps (spatial features) and wavelet decomposition (Doppler frequency features), as described in the previous sub-section. The network utilizes the encoder component of a transformer architecture to convert these features into a sequence of embeddings. Each embedding encapsulates both the information of its corresponding input feature and its contextual relationships with other features in the sequence. Firstly, positional embedding gives the model information about the order of the features, such as the position of the current feature, and the dependencies with each other input features. Here, the trainable embedding method is used, allowing the network to learn the positional relationships directly through additional parameters. Secondly, the embeddings and features are passed to a multi-head (4-head) self-attention (MHSA) block. The MHSA allows the model to focus on different parts of the input features simultaneously. The embeddings are projected into queries (Q), keys (K) and values (V) matrices. Attention scores are given by the following equation:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}V\right) \tag{3}$$

where d_k is the dimension of the key vectors and helps normalize the dot product value. Thirdly, an LSTM layer enables the model to retain past information, helping to handle sequential input data, such as features extracted across multiple data frames. Following the LSTM layer, a global max pooling layer is applied to extract salient features according to the input sequence. Finally, since the task is formulated as a classification problem, two fully connected (FC) layers are used to reduce dimensionality and produce class scores, followed by a softmax layer for final prediction. The network outputs a $1 \times k$ vector representing the probability distribution over the number of grouped target categories. The classification results will be passed to the tracker to compute the OSPA metric which also include positional, location-based error.

For the training phase, the Adam optimizer and cross-entropy loss function are used for backpropagation. The learning rate is set to 0.001, the batch size is 64, and the model is trained for 20 epochs.

(a) Experimental room

(b) Setup of INRAS radar and Azure Kinect DK camera

Fig. 4. Pictures of the radar laboratory environment for data collection in TU Delft and of the radar with Kinect camera.

III. Experimental Setup and Results analysis

A. Experimental Setup

In this work, a commercial 24GHz FMCW radar (by Joby Austria, former INRAS) with a relatively narrow bandwidth of 250MHz is used to validate the performance of the proposed approach. The relatively small bandwidth results in a range resolution of approximately 60 cm, which indicates that the target occupies a large area in the range profile while also making it difficult to isolate individual body parts. The detailed parameters used are listed in Table I. The radar is equipped with 15 virtual channels in azimuth which provide a fine angular resolution.

To evaluate the accuracy of multi-target tracking systems, the OSPA [8] metric is considered. From the following equation, the OSPA metric has two components: one is the distance error (related to localization performance), and the other one is the cardinality error (related to the number of people recognized to be in the room.

$$OSPA = (d_{loc}^{p} + d_{card}^{p})^{1/p} \tag{4}$$

where p denotes the order, which is set to 2 in this case. In order to include the classification error in the metric, the cardinality component is modified as follows, i.e., $d_{card} = \{\frac{(n+q)-m}{n+q} c^p\}^{1/p}$, where n is number of considered tracks, m is the number of ground truth tracks, and q is the difference between classifier predictions and true number of targets present in the scene.

For the validation of the algorithm, data collection was performed in a laboratory room of the MS3 group at TU Delft. The campaign involved 5 individuals, and up to 3 persons move simultaneously as a group in the room. The environment is shown in Fig. 4a. An auxiliary RGBD camera (Azure Kinect DK [17]) was used to collect ground truth data at the same time as the radar measurements. The RGBD data was processed by [18], with the center of the bounding box considered as the ground truth human position.

In total, 6 different movement scenarios are performed and collected. These scenarios include a mixture of different walking patterns with variable distances between the people in case of groups of them being present (Table II). Most of the activities are recorded for around 20 minutes. This leads to having around 51000 frames of data collected and used for validation.

The feature values for a single time-stamp of the sequences are extracted by a sliding window of 3 seconds to capture Doppler information with a proper resolution. In order to prepare the sequence of features in the correct format for the subsequent network, feature values extracted from consecutive time-stamps are stacked and an input matrix is formed. For each activity, the testing data is separated by selecting the first portion of each sequence for an amount equivalent to 15% of feature samples. The rest of the feature samples are used for training (60%) and validation (25%).

TABLE II
Summary of the 6 different movement scenarios collected for testing the proposed method.

#NO.	Movement Scenarios	Duration (min)
①	1 Target walking forward and backward	20
②	1 Target randomly walking	10
③	2 Targets walking forward and backward	20
④	2 Targets walking and following each other	5
⑤	2 Targets randomly walking	10
⑥	3 Targets walking forward and backward	20

TABLE III
Testing accuracy for different numbers of frame features as an input.

Number of Frames stacked	Accuracy (%)
2	91.89
5	89.61
10	92.75
15	92.35
20	91.73

B. Analysis of length of stacked features

The features stacked with varying time lengths (i.e., across a different number of data frames) can capture different information, particularly in the Doppler domain. Table III presents the testing accuracy obtained using Doppler features from all levels of the wavelet decomposition combined with spatial features. The highest testing accuracy of 92.75% is achieved when features from the current frame are stacked with those from the previous 9 frames. However, accuracy begins to decline when more than 10 frames are included. Notably, stacking with only 5 frames results in the lowest accuracy among the tested configurations.

C. Analysis of features

Unlike conventional wavelet-based signal decomposition, the MODWT approach maintains the same data size for each level and also keeps information redundancy between each level. The corresponding frequency bands of each level is listed in Table V. Generally, the Doppler frequencies associated with human walking signatures appear predominantly in Level 2 and Level 3. When a target changes direction, the frequency components takes lower values, approaching the 'approximation' level (which covers between 0 $\sim$ 28Hz). When the network is trained using frequency domain features alone, Level 3 yields the best performance among the five levels, highlighting its importance. However, using frequency features alone proves insufficient for achieving

TABLE IV
SUMMARY OF PERFORMANCE METRICS FOR TRACKING WITH PROPOSED NETWORK CLASSIFIER.

Scenarios	Proposed			SVM-based counting [6]		
	ACC (%)	RMSE (cm)	OSPA	ACC (%)	RMSE (cm)	OSPA
① 1 target walking	99.88	20.64	0.25	95.56	20.51	0.27
② 1 target random walking	86.15	28.66	0.37	79.46	27.77	0.41
③ 2 targets walking	96.68	22.53	0.23	100	22.68	0.21
④ 2 targets following	90.44	40.48	0.48	89.71	40.06	0.5
⑤ 2 targets random walking	94.14	35.62	0.39	83.86	35.15	0.43
⑥ 3 targets walking	87.68	29.27	0.32	84.58	29.34	0.34
Averaged	**92.50**	29.53	**0.34**	88.86	29.25	0.36

TABLE V
ACCURACY COMPARISON FOR COMBINATIONS OF DIFFERENT LEVELS OF FEATURES. (*Approx* DENOTES THE LOWEST FREQUENCY FROM $0 \sim 28$HZ.)

MODWT Level	Frequency (Hz)	ACC of frequency features (%)	ACC of Spatial + Frequency (%)
Approx	0-28	75.94	90.68
level4	27-58	72.31	89.33
level3	54-116	77.16	89.98
level2	109-233	75.9	88.58
level1	225-450	74.21	89.40

TABLE VI
ACCURACY COMPARISON OF DIFFERENT NETWORKS AND CONVENTIONAL METHOD.

	ACC (%)	Averaged OSPA
Proposed	92.50	0.340
SVM [6]	88.86	0.360
LSTM [20]	89.42	0.353
BiLSTM [21]	86.56	0.362
GRU [22]	89.13	0.352
1D TCN [19]	87.49	0.360

optimal accuracy. Therefore, combinations of spatial features with individual frequency levels were evaluated. All such combinations achieve accuracy close to 90%, with relatively small differences between levels. Based on Doppler frequency characteristics and classification performance, the combination of spatial features with Level 3 frequency features is selected as the most promising result for a final validation.

D. Analysis of overall tracking & counting performance

The overall tracking and individual counting performance are summarized in Table IV. The proposed method achieves approximately 4% higher classification accuracy on the testing data compared to the more conventional SVM-based approach. Notably, in the random walking scenario, it outperforms the baseline by at least 7%, demonstrating robustness to frequent direction changes that typically degrade Doppler-based features. This improvement shows the ability of the network to capture significant temporal relationships across frames. The only scenario where performance drops about 4% is in the case of two targets walking simultaneously. However, the averaged RMSE difference is approximately equivalent to 0.25 centimeters, which could be considered as not too significant. Overall, the proposed method also achieves better performance in terms of the averaged OSPA metric compared to the SVM classifier.

E. Analysis of different neural networks and conventional method

Results for conventional methods are compared in Table VI. Firstly, the conventional statistical method (i.e., based on SVM classifier) used in previous research [6] shows 88.86% accuracy and 0.360 OSPA error. Then, the remaining three methods compared are different network architectures that can handle time series data and have memory units. The LSTM and Gated Recurrent Unit (GRU) shows similar OSPA error around a value of 0.35. The bidirectional LSTM does not appear to work well in this specific classification task. Finally, a 1D Temporal Convolutional Network (TCN) from [19] is implemented to handle the sequence features with a resulting accuracy of 87.49%. It should be noted that in this study conventional Convolutional Neural Networks (CNN) are not considered because they would need a 2D, image-like input instead of the sequence-like data series considered here. In summary, the proposed method with a transformer-based architecture shows both the highest accuracy and the lowest OSPA error.

IV. CONCLUSION

In this paper, a transformer-based network is proposed and designed to count the number of people in a group while they are in motion and tracked. The pipeline is designed to handle the people moving independently and changing their direction of motion. The proposed approach combines the tracking framework with the network classifier to address the counting problem while simultaneously tracking. The approach is implemented and validated using a 24GHz MIMO FMCW radar. As a result, the proposed method demonstrate a classification accuracy of 92.5% and an OSPA tracking metric of 0.34, outperforming a previously proposed SVM classifier operating on the same data.

ACKNOWLEDGMENT

This research was in part financially supported by Huawei Sweden Gothenburg Research Center. The authors are grateful to the volunteers who participated in the data collection.

REFERENCES

[1] Texas Instruments, "PeopleTrackingandCounting Reference Design Using mmWave Radar Sensor.pdf." [Online]. Available: https://www.ti.com/lit/ug/tidue71d/tidue71d.pdf

[2] A. Ninos, J. Hasch, M. Heizmann, and T. Zwick, "Radar-Based Robust People Tracking and Consumer Applications," *IEEE Sensors Journal*, vol. 22, no. 4, pp. 3726–3735, Feb. 2022.

[3] D. Wang, J. Park, H.-J. Kim, K. Lee, and S. H. Cho, "Noncontact Extraction of Biomechanical Parameters in Gait Analysis Using a Multi-Input and Multi-Output Radar Sensor," *IEEE Access*, vol. 9, pp. 138 496–138 508, 2021.

[4] L. Ren, A. Yarovoy, and F. Fioranelli, "Grouped People Counting Using mm-wave FMCW MIMO Radar," *IEEE Internet of Things Journal*, pp. 1–1, 2023.

[5] N. Knudde, B. Vandersmissen, K. Parashar, I. Couckuyt, A. Jalalvand, A. Bourdoux, W. De Neve, and T. Dhaene, "Indoor tracking of multiple persons with a 77 ghz mimo fmcw radar," in *2017 European Radar Conference (EURAD)*. IEEE, 2017, pp. 61–64.

[6] D. Wang, S. Yuan, A. Yarovoy, and F. Fioranelli, "Grouped target tracking and seamless people counting with a 24 ghz mimo fmcw," 2025, under review at IEEE Transactions on Radar Systems. [Online]. Available: https://arxiv.org/abs/2504.04969

[7] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, Jul. 1998.

[8] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A Consistent Metric for Performance Evaluation of Multi-Object Filters," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3447–3457, Aug. 2008.

[9] D. Wang, F. Fioranelli, and A. Yarovoy, "Analysis of Processing Pipelines for Indoor Human Tracking Using FMCW Radar," in *2024 IEEE Radar Conference (RadarConf24)*, May 2024, pp. 1–6.

[10] ——, "Quantitative Assessment of People Tracking with FMCW MIMO Radar," in *2024 21st European Radar Conference (EuRAD)*, Sep. 2024, pp. 380–383.

[11] F. Xiao, T. Lu, M. Wu, and Q. Ai, "Maximal overlap discrete wavelet transform and deep learning for robust denoising and detection of power quality disturbance," *IET Generation, Transmission & Distribution*, vol. 14, no. 1, pp. 140–147, 2020.

[12] J. Quilty and J. Adamowski, "A maximal overlap discrete wavelet packet transform integrated approach for rainfall forecasting – A case study in the Awash River Basin (Ethiopia)," *Environmental Modelling & Software*, vol. 144, p. 105119, Oct. 2021.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Aug. 2023.

[14] 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," Jun. 2021.

[15] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," Mar. 2021.

[16] K. Cao, T. Zhang, and J. Huang, "Advanced hybrid lstm-transformer architecture for real-time multi-task prediction in engineering systems," *Scientific Reports*, vol. 14, no. 1, p. 4890, 2024.

[17] "Azure Kinect DK – Develop AI Models | Microsoft Azure." [Online]. Available: https://azure.microsoft.com/en-us/products/kinect-dk

[18] Yuxin Wu and Alexander Kirillov and Francisco Massa and and Wan-Yen Lo and Ross Girshick, "Detectron2," Meta Research, Feb. 2025. [Online]. Available: https://github.com/facebookresearch/detectron2

[19] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[22] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014, pp. 1724–1734.