

Article

Vision Transformers (ViT) for Blanket-Penetrating Sleep Posture Recognition Using a Triple Ultra-Wideband (UWB) Radar System

Derek Ka-Hei Lai ^{1,†}, Zi-Han Yu ^{2,†}, Tommy Yau-Nam Leung ¹, Hyo-Jung Lim ¹, Andy Yiu-Chau Tam ¹ , Bryan Pak-Hei So ¹, Ye-Jiao Mao ¹, Daphne Sze Ki Cheung ³ , Duo Wai-Chi Wong ^{1,*}  and James Chung-Wai Cheung ^{1,4,*} 

¹ Department of Biomedical Engineering, Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China

² School of Engineering Sciences, Huazhong University of Science and Technology, Wuhan 430074, China

³ School of Nursing, The Hong Kong Polytechnic University, Hong Kong 999077, China

⁴ Research Institute of Smart Ageing, The Hong Kong Polytechnic University, Hong Kong 999077, China

* Correspondence: duo.wong@polyu.edu.hk (D.W.-C.W.); james.chungwai.cheung@polyu.edu.hk (J.C.-W.C.); Tel.: +852-2766-7669 (D.W.-C.W.); +852-2766-4534 (J.C.-W.C.)

† These authors contributed equally to this work.

Abstract: Sleep posture has a crucial impact on the incidence and severity of obstructive sleep apnea (OSA). Therefore, the surveillance and recognition of sleep postures could facilitate the assessment of OSA. The existing contact-based systems might interfere with sleeping, while camera-based systems introduce privacy concerns. Radar-based systems might overcome these challenges, especially when individuals are covered with blankets. The aim of this research is to develop a nonobstructive multiple ultra-wideband radar sleep posture recognition system based on machine learning models. We evaluated three single-radar configurations (top, side, and head), three dual-radar configurations (top + side, top + head, and side + head), and one tri-radar configuration (top + side + head), in addition to machine learning models, including CNN-based networks (ResNet50, DenseNet121, and EfficientNetV2) and vision transformer-based networks (traditional vision transformer and Swin Transformer V2). Thirty participants ($n = 30$) were invited to perform four recumbent postures (supine, left side-lying, right side-lying, and prone). Data from eighteen participants were randomly chosen for model training, another six participants' data ($n = 6$) for model validation, and the remaining six participants' data ($n = 6$) for model testing. The Swin Transformer with side and head radar configuration achieved the highest prediction accuracy (0.808). Future research may consider the application of the synthetic aperture radar technique.

Keywords: ablation study; deep learning; feature extraction; sleep monitoring; obstructive sleep apnea



Citation: Lai, D.K.-H.; Yu, Z.-H.; Leung, T.Y.-N.; Lim, H.-J.; Tam, A.Y.-C.; So, B.P.-H.; Mao, Y.-J.; Cheung, D.S.K.; Wong, D.W.-C.; Cheung, J.C.-W. Vision Transformers (ViT) for Blanket-Penetrating Sleep Posture Recognition Using a Triple Ultra-Wideband (UWB) Radar System. *Sensors* **2023**, *23*, 2475. <https://doi.org/10.3390/s23052475>

Academic Editors: Anastasios Doulamis, Nikolaos Doulamis and Athanasios Voulodimos

Received: 19 January 2023

Revised: 16 February 2023

Accepted: 20 February 2023

Published: 23 February 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Obstructive sleep apnea (OSA) is one of the most common sleep breathing disorders, with a prevalence of 9% to 38% that increases with age [1]. Untreated OSA patients may stop breathing numerous times every night when they sleep [2]. In order to “restart” breathing, the brain awakes, which leads to poor and fragmented sleep [2]. Sleep apnea has serious health repercussions and elevates the risk of diabetes, heart disease, hypertension, and heart failure if left untreated [3]. The care of concomitant neurological diseases, such as epilepsy, stroke, multiple sclerosis, and headache also becomes burdensome [4]. OSA has caused an economic burden of USD 6.5 billion in accidents and injuries, USD 86.9 billion in lost productivity at work, and USD 30 billion in healthcare annually in the USA [5].

There is an established relationship between OSA and sleep positions/postures [6]. A supine posture might significantly reduce the risks of OSA because it prevents the prolapse

of the tongue and the soft palate against the pharyngeal wall by gravity [6,7]. In addition, a prone position presses on the lungs and affects respiration [7]. Contrarily, a lateral position (or side-lying posture) resolves the issue by maintaining the retropalatal and retroglossal airways [8]. These findings also support the crucial impact of the sleep position on the incidence and severity of sleep apnea [6–8]. In order to assess the sleep positions of OSA patients and their rehabilitation progress, sleep posture recognition and tracking could be one of the essential assessment components [9].

Various sensors have been developed to monitor sleep postures and behaviors, including body pressure sensors, physiological sensors, cameras (and depth cameras), and wearable devices [10]. The pressure intensity distribution generated by a pressure mat has been utilized to characterize sleep postural behavior and estimated sleep quality [11–13]. Video recordings using red–green–blue (RGB) or red–green–blue–depth (RGB-D) images can capture and facilitate observation of the sleep postures of individuals directly [14–17]. Wearable devices using actigraphy or accelerometry can measure physical activity and infer motor or behavioral activities [18]. Therefore, spectrogram analysis of data from wearable devices can be used to estimate sleep postures via the movement of body segments [19]. However, these systems or sensors can be expensive or interfere with sleep, which discourages practical use. Optical sensors or cameras suffer from interference from ambient light sources [20], in addition to privacy concerns [21].

Radar-based techniques might overcome these challenges and have demonstrated applications to sleep posture recognition [22]. In fact, there are different kinds of radar signals. Continuous wave radar is the most common type that sends and receives frequency signals continually, but it has poor discriminability with pulse and respiratory signals [22]. Frequency-modulated continuous wave (FMCW) radar cannot remove the weakness and is vulnerable to radio incoherence from unsteady object motion and micro-Doppler signals [23]. Impulse radio ultra-wideband radar (IR-UWB) can detect multiple objects and evaluate distance with less radiation [23], which facilitates applications, such as in body movement analyzers and trackers [24,25]. Ahmed and Cho [26] analyzed the waveform of IR-UWB to differentiate hand movements and gestures. Rana et al. [27] developed a markerless gait analysis system using IR-UWB technology. Recently, Lai et al. [28] attempted to classify sleep postures using IR-UWB signals by identifying their key statistical features.

The posture recognition function is often facilitated by machine learning techniques, especially tree-based and convolutional neural network (CNN) models. Piriyaikitakonkij et al. [29] designed a CNN model, SleepPoseNet, that applied a feature-mapped matrix of time and frequency domains to estimate transitional postures. Kiriazi et al. [30] applied the decision tree to the effective radar cross section (ERCS) and displacement magnitude information to distinguish stationary torso postures. Zhou et al. [31] transformed a radar signal to image features, which were then handled by a CNN model integrated with an inception-residual module. Tam et al. [14] guided a CNN-based deep learning model (ECA-Net) by generating anatomical landmarks on depth images using a pose estimator. In addition, using random forest, Lai et al. [28] extracted radar features from each radar bin range for posture recognition. Although CNN models are widely used because of their capability to understand higher level semantic features and their superior performance [32], they were poor in understanding the global representation that might affect the performance of sleep posture recognition.

A branch of deep learning models, vision transformers (ViTs), has emerged recently [33–35]. The origin of ViT, “Transformer”, was designed for natural language processing (NLP) and was later applied to visual computing tasks, such as object detection [36] and segmentation [37], and human motion recognition [38], such as pose estimation [39,40], and gait recognition [41,42]. The “Transformer” built upon the sequence-to-sequence encoder–decoder architecture and substituted the recurrent layers with attention mechanism, enabling the long-term memory of every token (word) [43]. While CNN applied pixel arrays in the model and lost spatial relationship information in the pooling layers [44], ViT has a substantially different backbone and model architecture from CNN models. It

embeds and segments images into small patches followed by a self-attention mechanism without the use of convolutional layers [34]. The patches and positional embedding are input to the transformer encoder, which originally operates on tokens (words) [34]. ViT has a higher computational efficacy and accuracy than CNN models but requires more training data [45,46].

Our study was motivated by the need for sleep posture assessment for OSA patients that might not be practically fulfilled by the current systems because of cost, privacy concerns, and the interference with sleep. The novelty of this study lies in the transformation of multiple radar signals to a spatiotemporal graph that can be input to the cutting-edge deep learning model, ViT, for sleep posture recognition. As we were interested in different configurations of the radar systems, the radars were placed on the ceiling (top radar), at the side of the participant (side radar), and on top of the head of the participants (head radar), as shown in Figure 1. We assumed that the tri-radar configuration (top + side + head) could improve the accuracy of posture prediction as compared to the single radar (top, side, and head) and the dual-radar configurations (top + side, top + head, and side + head). In addition, we compared different deep learning models, including CNN-based models (ResNet [47], DenseNet [48], and EfficientNet [49]) and ViT (traditional ViT [34] and Swin vision transformer [50]). We hypothesized that vision transfers with the tri-radar configuration (top + side + head) would outperform the others.

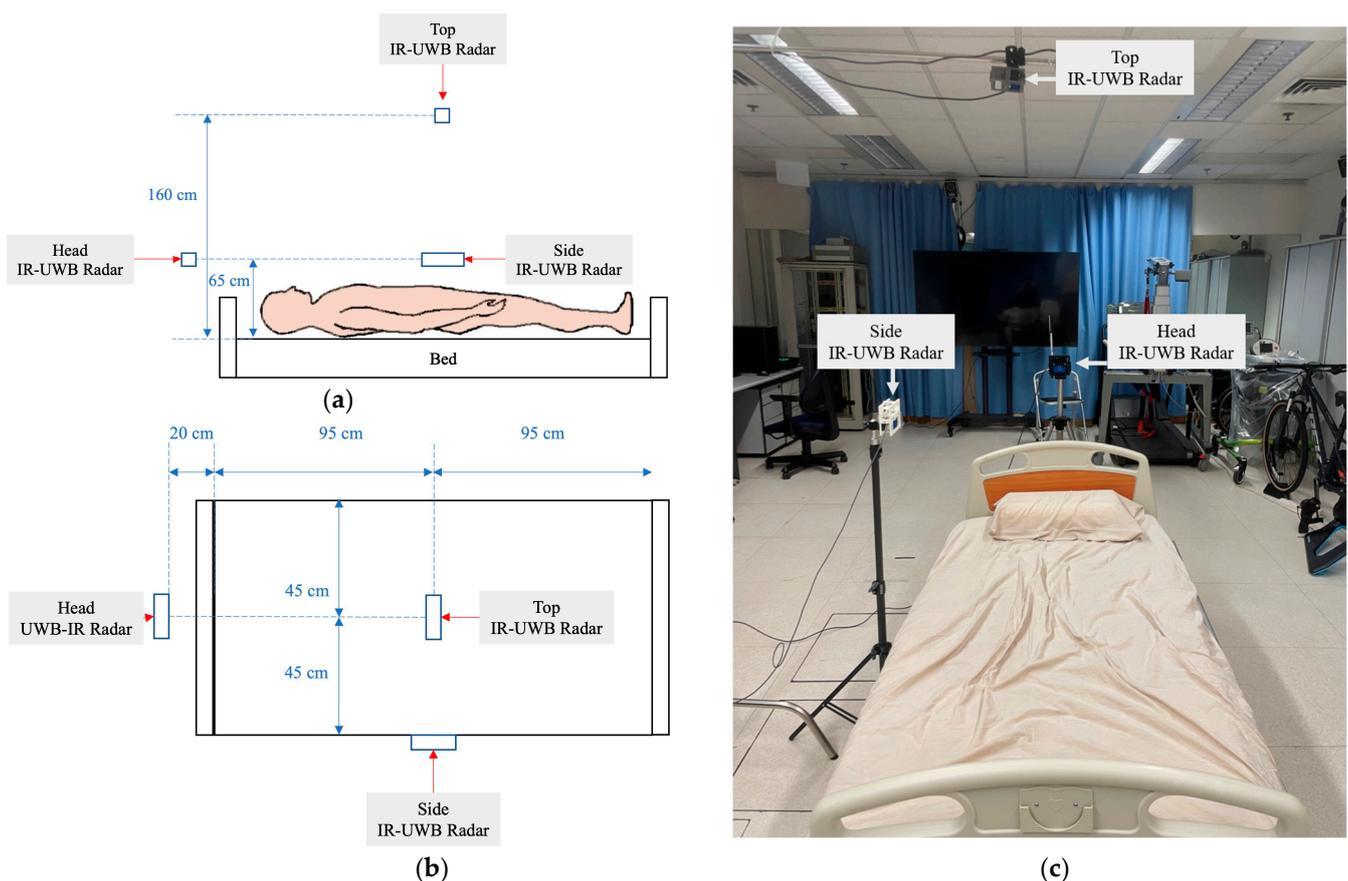


Figure 1. Schematic diagram of the system setup in (a) side view; (b) top view; and (c) photo.

2. Materials and Methods

2.1. Hardware and Software Configuration

Three IR-UWB radar sensor system-on-chips (Xethru X4M03 v5, Novelda, Oslo, Norway) were used. Each sensor consisted of a fully programmable system controller and an antenna. The transmitter center frequency and energy per pulse were 7.29 GHz and 2.65 picojoules, respectively, which complied with the ETSI/FCC. The receiver had a sam-

pling rate of 23.328 GS/s, a total radar frame length of 9.87 m, and the distance between each radar bin was 0.00643 m. The receiver gain noise figures were 14.1 dB and 6.7 dB, which also met the ETSI/FCC compliance requirement. Both the range of elevation angle and the azimuth angle were between -65° and $+65^\circ$. The other parameters are shown in Table 1. The detection range was adjusted to encompass the region of interest (RoI).

Table 1. Configurations of the IR-UWB radar devices.

Parameters	Top Radar	Side Radar	Head Radar
Detection Range	1.3 m–1.8 m	0.6 m–1.1 m	0.4 m–1.2 m
Transmission Power	6.3 dBm	6.3 dBm	6.3 dBm
Pulse Repetition Frequency	15.188 MHz	15.188 MHz	15.188 MHz
Bin Resolution	78 bins	78 bins	125 bins
Frame Rate	20 frames/second	20 frames/second	20 frames/second

2.2. System Setup

The system setup involved three IR-UWB radars with associated connection cables, a height adjustable hospital bed, a comforter, two tripods, one light boom, and a laptop computer. As shown in Figure 1, the bed was 0.5 m from the ground. The top radar was hung on a light boom, which was 1.5 m from the bed surface close to the height of a household ceiling. The side radar was mounted 0.65 m from the bed surface using a tripod, at the longer edge of the bed, according to the minimum detection distance requirement of the IR-UWB radar. The head radar was mounted 0.65 m from the bed surface and 0.2 m from the shorter edge using another tripod, to accommodate the effective angle of the depression (60°) of the radar. The three sensors were positioned orthogonally over the hospital bed (1.9 m \times 0.9 m \times 0.5 m with mat).

2.3. Participant Recruitment and Data Collection

All recruited participants signed an informed consent after receiving an oral and written description of the experiment before beginning the experiment, which was approved by the Institutional Review Board (reference number: HSEARS20210127007). The inclusion criteria included healthy adults aged over 18. In this study, we recruited 30 young adults (19 males and 11 females). Their average age was 22 (SD: 2.00, range 18–27). The mean weight and height were 170 cm (SD: 9.64 cm, range 156.5–196 cm) and 62.5 kg (SD: 12.70 kg, range 46–100 kg), respectively. The exclusion criteria included physical disability, obesity, pregnancy, or any cardiorespiratory problems, in addition to participants with difficulties in maintaining or switching specific postures in bed.

Before the experiment, participants were instructed to remove clothing or accessories with metallic components (such as a belt with a metallic buckle), in addition to their shoes and outerwear. Throughout the experiment, they were asked to lie on the bed with a support pillow, covered by a comforter. They were then instructed to lie in different postures, in the order of (1) supine, (2) right lateral, (3) left lateral, and (4) prone, as shown in Figure 2. A ringing bell was played to notify the participants to adopt their assigned posture with self-chosen comfortable limb placement. After the participants finalized their posture, we then started the recording. Each posture was recorded for 15 s. The full course was repeated ten times. We collected 1200 samples (30 participants \times 4 postures \times 10 repetitions). The samples were labelled manually during the experiment.

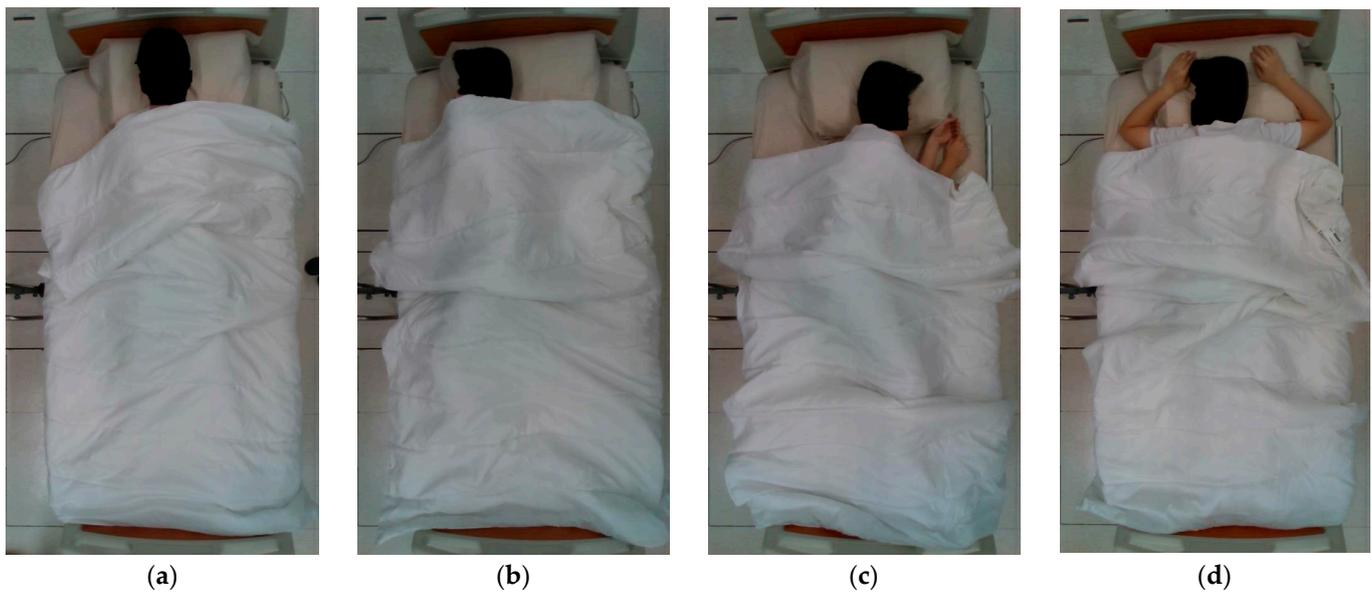


Figure 2. The four under-blanket recumbent postures: (a) supine; (b) right side-lying; (c) left side-lying; (d) prone.

2.4. Data Processing

The data processing pipeline comprised preprocessing, denoising, augmentation, resizing, and merging. The top and side radars produced 78 bins per frame, while the head radar produced 125 bin per frame. In the preprocessing stage, the last ten seconds of the recording were extracted for analysis. In the denoising stage, static objects in the scene were removed using the mean subtraction method by taking the average and subtracting it from each sample (Equation (1)). Then, background suppression was performed to remove the environmental noise (Equation (2)).

$$X'[n, m] = X[n, m] - \frac{1}{N} \sum_{i=0}^{N-1} X[n, i] \quad (1)$$

$$Y[n, m] = X'[n, m] - \frac{1}{M} \sum_{i=0}^{M-1} X[i, m] \quad (2)$$

where n was the fast time index (radar bin), and m was the slow time index (frame number).

Data augmentation techniques, including scaling, magnitude warping, flipping, time warping, and window warping [51,52] were applied (Figure 3). The scaling process multiplied each frame of the signal by a random scalar. For magnitude warping, the time series of each bin was multiplied by a curve generated using a cubic spline of 4 knots and $\sigma = 0.2$ [51,52]. The flipping process flipped at the center timepoint. Time warping perturbed the signal using the magnitude wrapping curve; while for window warping, a random window of 10% of the original duration wrapped the time dimension by 0.5 times or 2 times [51,52].

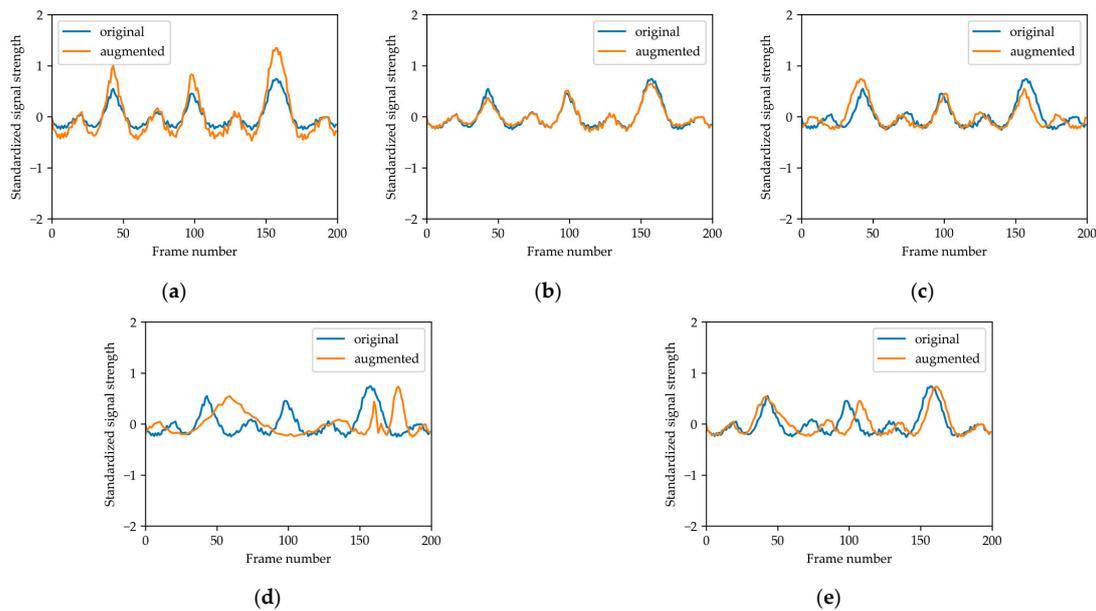


Figure 3. The five data augmentation strategies applied in our system, including: (a) scaling; (b) magnitude warping; (c) flipping; (d) time warping; and (e) window warping.

The augmented data were resized to images of $224 \text{ pixels} \times 224 \text{ pixels}$ to accommodate the input size requirements of the deep learning model (Figure 4). In the merging stage, the data of the three radars were distributed to three channels (RGB) to imitate an image. For the single-radar configuration, the data were cloned for the three channels. For the dual-radar configuration, the data of the first radar were input to the red channel, while those of second radar were input to the green channel, and a 224×224 zeroes array was assigned to the blue channel. For the tri-radar configuration, the top, side, and head radar corresponded to the red, green, and blue channels, respectively. Figure 4 illustrates the imitated image visualization of different radar configurations.

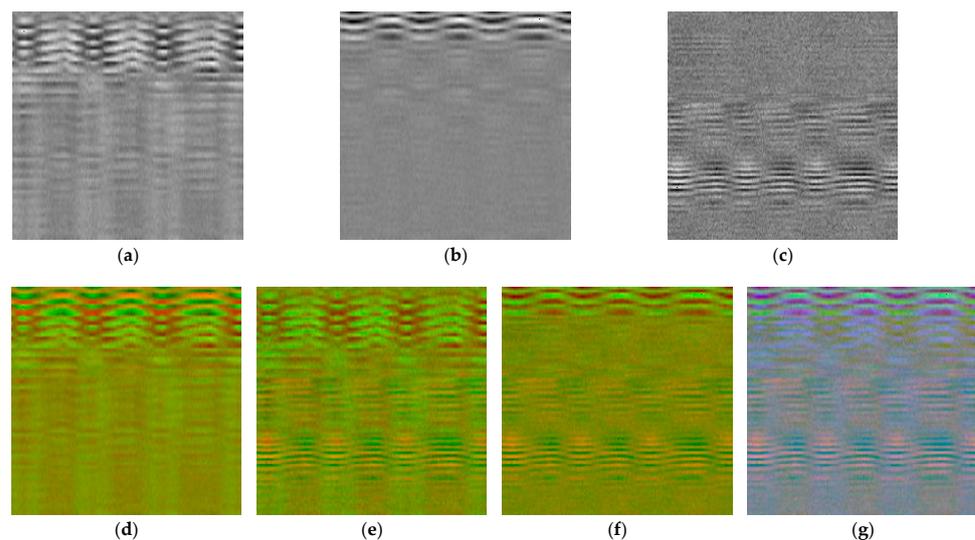


Figure 4. The imitated images generated by the single-radar configurations: (a) top radar, (b) side radar, and (c) head radar; the dual-radar configurations: (d) top + side radars, (e) top + head radars, (f) and side + head radars; and the tri-radar configurations: (g) top + side + head radars. The x -direction of the image represents the bin resolution, and the y -direction represents time. The resolution of the images was unified and resized to $224 \text{ pixels} \times 224 \text{ pixels}$ based on the data of the radar.

2.5. Model Training

The models were pretrained by ImageNet [53]. We trained the model using the data of 18 randomly selected participants, validated the model using the data of six participants, and tested it using the data of the other six participants. The performance from the three convolutional-based models (ResNet50, DenseNet121, and EfficientNetV2) and the two attention-based models (vision transformer and Swin Transformer V2) was compared in this study. The cross entropy loss was regarded as the loss function. The stochastic gradient descent using an initial learning rate of 0.001 and a momentum of 0.9 was applied as the optimizer. The learning rate was scaled down 10 times every 20 training epochs. Every model was trained for 100 iterations.

2.6. Model Validation

We used accuracy as the primary outcome of the models, which was defined by the fraction of correct predictions over the total number of predictions on the testing set. The validation dataset adjusted the model weights and attempted to minimize overfitting by facilitating early stopping. The accuracy was calculated by comparing the model prediction with the testing dataset.

3. Results

3.1. Performance of Different Models

The transformer-based models performed generally better than the convolutional-based model, where the average accuracies were 0.613 and 0.637 for the vision transformer and Swin Transformer, respectively, compared to 0.551, 0.543, and 0.538 for the ResNet50, DenseNet121, and EfficientNetV2, respectively, as shown in Figure 5. Among all models, the Swin Transformer with the side + head radar configuration produced the best prediction accuracy (0.808).

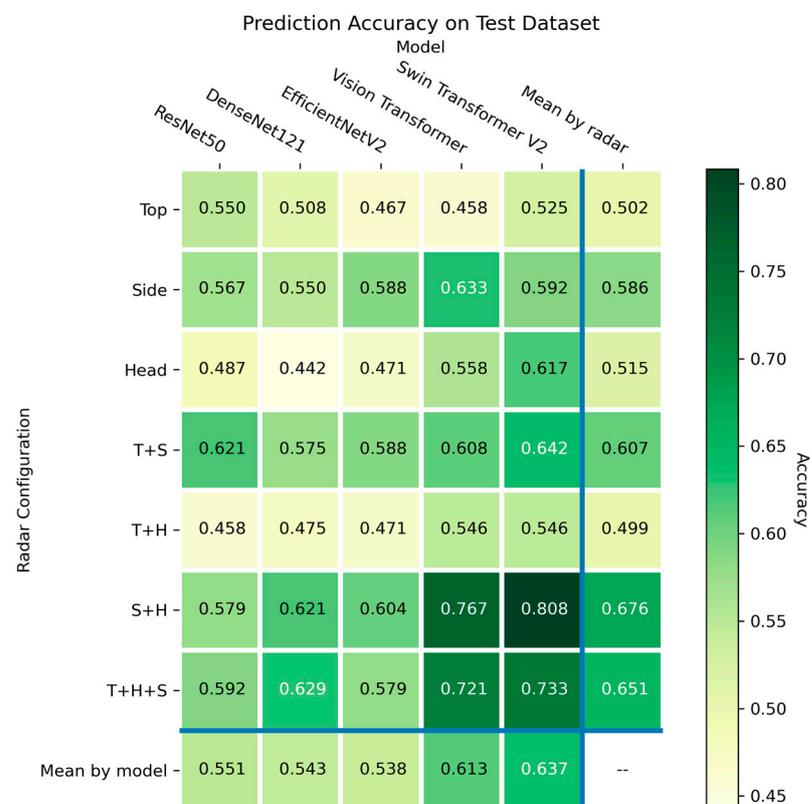


Figure 5. Heatmap showing the prediction accuracy of the different machine learning models and radar configurations. H: head radar; S: side radar; T: top radar.

3.2. Performance of the Radar Configurations

Overall, the dual-radar was able to produce the better result, followed by the tri-radar and the single-radar. From Figure 5, the average accuracies of the dual-radar configurations were 0.676, 0.607, and 0.499 for the side + head, top + side, and top + head configurations, respectively, compared to 0.651 for the tri-radar configuration, and 0.586, 0.515, and 0.502 for the side, head, and top configuration, respectively. Among all configurations, the side + head configuration with Swin Transformer yielded the best accuracy (0.808).

3.3. Subgroup Analysis on Posture Conditions

We extracted the confusion matrix of the Swin Transformer with the side + head radar configuration, which had the best prediction outcome (Figure 6). For a total of 120 predictions, the side-lying postures had 110 correct predictions, while the supine/prone had 103 predictions. The supine postures gave the largest number of correct predictions (55/60), followed by right side-lying (54/60), left side-lying (49/60), and prone (36/60).

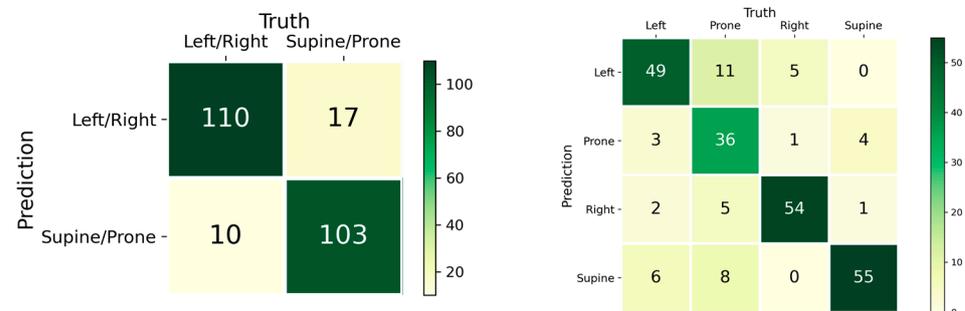


Figure 6. Subgroup analysis of the posture conditions using a confusion matrix over the prediction performance of the Swin Transformer with dual-radar configuration (side + head).

4. Discussion

The novelty of this study lies in the application of the vision transformer deep learning models and multiple radar configurations to improve the sleep posture classification accuracy. The challenges of the radar system were that it could not effectively distinguish a stationary target from clutter, though existing radar processing techniques have mainly focused on moving target detection [54]. In addition, radar processing often relies on frequency analysis, such as fast Fourier transform (FFT), which could be more sensitive to dynamics and biomotions (respiration and heartbeat) but relatively insensitive to stationary postures [54]. We addressed these challenges by applying multiple radar systems that could reflect the different body cross-sectional areas in different postures.

Both the convolutional-based and transformer-based models were capable of extracting the pattern, but the transformers could further facilitate positional encoding [55]. In particular, this mechanism allowed the transformer models to locate the fringes, which was an important feature to distinguish different postures. Among the two transformer models, the Swin Transformer utilized shifted window and masked signal modeling techniques. On the other hand, the CNN models, such as EfficientNet, used the compound scaling method, by optimizing the network depth, network width, and input image resolution [56]. Nevertheless, the imitated images had low resolution, and the potential of EfficientNet could not be unleashed. In contrast, ResNet and DenseNet were designed for an input image with lower resolution; their higher performance over EfficientNet reflected that the two models might be more suitable in our image resolution setting. ResNet and DenseNet differ in how they connect the feature maps. ResNet sums all feature maps, while DenseNet concatenates them [57]. In our study, ResNet had superior performance on single radar settings, but DenseNet was better on dual-radar and tri-radar settings, showing that the summation approach might work better on single radar settings, and concatenation might work better on dual-radar and tri-radar settings.

Our results indicated that the configurations involving top radars produced the worse results. The top radar had a distant placement, whereas it covered largest RoI of human body, resulting in the highest electric field attenuation (energy loss). This produced a poor signal-to-noise ratio (SNR) that affected the prediction accuracy. In addition, we distributed the radio-frequency data into three channels to imitate RGB images, where the noise from the top radar might intervene with the latent feature mining from other radars. A low SNR could worsen the prediction, both in the single radar and dual-radar configurations. Enhancing the signal-to-noise ratio for the radar could be one method to improve classification, which could be achieved by methods such as high order cumulants (HOCs), ensemble empirical mode decomposition (EEMD), complex signal demodulation (CSD), and the state space method (SSM) [58]. Among all these methods, Liang, et al. [58] suggested that CSD worked better for IR-UWB usage. We could further enhance the radar signal with the complex signal demodulation (CSD) algorithm.

Non-lateral (prone and supine) postures aggravate OSA, while lateral (left or right) postures do not induce a negative impact on OSA. In our study, we presented a subgroup analysis graph over the lateral and non-lateral predictions. The best model achieved 213 (or 0.888 in percentage) correct prediction counts. We believe that this system could utilize this information to alert OSA patients to maintain lateral postures to mitigate the problem.

The accuracy of our study and related recent studies was compared, as shown in Table 2. The accuracy of the existing systems ranged from 0.737 [29] to 0.984 [30], while our system had an accuracy of 0.808 using the state-of-the-art vision transformer model. Studies inputting handcrafted features on machine learning models (i.e., no deep learning) demonstrated better results. The sample size might affect the model performance [59], and the size of the existing studies varied from 8 to 120, with different numbers of classified postures (from 3 to 12). In addition, a covering blanket or quilt would introduce challenges to the model predictions.

Table 2. Comparison of the sleep posture accuracy between this study and other related recent studies.

Source	<i>n</i>	<i>np</i>	Stationary or Transitional	Hardware	Classifier	DL	Blanket	Accuracy
This study	30	4	Stationary	IR-UWB (Xethru X4M03)	Swin Transformer SleepPoseNet	Y	Y	0.808
Piriyajitakonkij et al. [29]	38	4	Transitional	IR-UWB (Xethru X4M03)	(Deep CNN with Multi-View Learning)	Y	N	0.737
Lai et al. [28]	18	4	Stationary	IR-UWB (Xethru X4M03)	Random Forest	N	N	0.938
Kiriazzi et al. [30]	20	3	Stationary	Dual frequency Doppler radar system	Decision Tree	N	N	0.984
Zhou et al. [31]	8	8	Transitional	FMCW radar system	CNN with Inception Residual module	Y	N	0.872
Tam et al. [14]	120	7	Stationary	Depth camera (Realsense D435i)	ECA-Net50	Y	Y	0.915
Mohammadi et al. [60]	12	12	Stationary	Depth camera (Microsoft Kinect)	CNN	Y	Y	0.760

CNN: convolutional neural network; DL: deep learning; FMCW: frequency-modulated continuous-wave; IR-UWB: impulse-radio ultra-wideband; *n*: sample size; *np*: number of postures; N: No; No.: number; Y: Yes.

There were some limitations in our study. Our proposal aimed to identify the best radar configurations that could collect the most representative latent information on sleep posture for prediction. Nonetheless, the signals collected for each radar were collapsed into a one dimensional time-series. In order words, we could not extract the complete spatial information of the body topology. The full geometrical information of the body posture would not only improve the accuracy of sleep posture prediction but also provide explainable information for the prediction. Future improvements may consider the applica-

tions of synthetic aperture radar techniques to obtain complete spatial information through scanning and sweeping the RoI periodically [61].

In addition, a large dataset is imperative for training machine learning models, especially deep learning models [62]. In this study, we recruited 30 participants with 10 repetitions of each posture and applied data augmentation techniques. More participants with different ages, sexes, and body builds could enhance the robustness and generalizability of the prediction. Age and sex might affect the preference of sleep posture and position, but they might not confound our model since we believe that they are not associated with the body and limb position of a posture. Nevertheless, body build (or body mass index) might have an impact on our model, since it attenuates the effective area for the radar signal reflection. On the other hand, model training and prediction with fine-grained posture classes on upper limb placements could be conducted. Some participants that put their hands on the front of their chest might cause interference with the measurement of the vital sign signals by the radar, since the vital signs and the source of the vital signs are important inputs for posture estimation. Our system did not isolate the vital sign signal. Therefore, we do not know whether the ViT used the vital sign as a salient feature.

The dataset size influences the accuracy in transfer learning with deep learning models [62]. Compared to previous studies, we had a larger dataset size in addition to applying augmentation techniques, which improved the generalization of the deep learning model; however, a larger dataset size remains preferable. During the experiment, we repeated the postures 10 times but allowed freedom of limb placement, which facilitated the model generalization in terms of the postures. Some participants might place their limbs on the face or chest, which could weaken the signals from the vital signs.

The long-term objective of this research is to develop a comprehensive sleep surveillance system that can monitor sleep postures and behaviors. Our previous studies developed a depth camera system to monitor bed-exiting events [63,64] and to classify sleep postures in a field setting [14,15,28]. In the future, we will explore synthetic aperture radar and advanced modeling techniques, for instance, DensePose, which could estimate and map the human pixels of an RGB image to the 3D surface of the human body in real time [65,66].

5. Conclusions

Our study showed that the dual-radar configuration (side + head) with the Swin Transformer model could achieve the best sleep posture prediction accuracy of 0.808. Nevertheless, the limitations of this study included the limited data for model training in addition to the incomplete spatial information generated by the radar system. Future studies may consider a larger dataset and the application of synthetic aperture radar techniques.

Author Contributions: Conceptualization, D.W.-C.W. and J.C.-W.C.; Data curation, D.K.-H.L., Z.-H.Y. and T.Y.-N.L.; Formal analysis, D.K.-H.L., Z.-H.Y. and T.Y.-N.L.; Funding acquisition, D.W.-C.W. and J.C.-W.C.; Investigation, D.K.-H.L., Z.-H.Y. and T.Y.-N.L.; Methodology, D.W.-C.W. and J.C.-W.C.; Project administration, D.S.K.C. and J.C.-W.C.; Resources, D.S.K.C. and J.C.-W.C.; Software, D.K.-H.L. and A.Y.-C.T.; Supervision, D.W.-C.W. and J.C.-W.C.; Validation, H.-J.L., B.P.-H.S. and Y.-J.M.; Visualization, D.K.-H.L., Z.-H.Y. and A.Y.-C.T.; Writing—original draft, D.K.-H.L. and Y.-J.M.; Writing—review and editing, D.W.-C.W. and J.C.-W.C.; All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the General Research Fund from the Research Grants Council of Hong Kong, China (reference number: PolyU15223822), as well as the internal fund from the Research Institute for Smart Ageing (reference number: P0039001) and the Department of Biomedical Engineering (reference number: P0033913 and P0035896) at Hong Kong Polytechnic University.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of The Hong Kong Polytechnic University (Reference Number: HSEARS20210127007).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The program, model codes, and updates presented in this study are openly available from GitHub at <https://github.com/BME-AI-Lab/Vision-Transformers-for-Blanket-Penetrating-Sleep-Posture-Recognition> (accessed on 1 February 2023). The video/image dataset is not publicly available since it would disclose identity and violate confidentiality.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Senaratna, C.V.; Perret, J.L.; Lodge, C.J.; Lowe, A.J.; Campbell, B.E.; Matheson, M.C.; Hamilton, G.S.; Dharmage, S.C. Prevalence of obstructive sleep apnea in the general population: A systematic review. *Sleep Med. Rev.* **2017**, *34*, 70–81. [[CrossRef](#)] [[PubMed](#)]
2. Strollo Jr, P.J.; Rogers, R.M. Obstructive sleep apnea. *N. Engl. J. Med.* **1996**, *334*, 99–104. [[CrossRef](#)] [[PubMed](#)]
3. Caples, S.M.; Gami, A.S.; Somers, V.K. Obstructive sleep apnea. *Ann. Intern. Med.* **2005**, *142*, 187–197. [[CrossRef](#)] [[PubMed](#)]
4. Ho, M.L.; Brass, S.D. Obstructive Sleep Apnea. *Neurol. Int.* **2011**, *3*, e15. [[CrossRef](#)] [[PubMed](#)]
5. Wickwire, E.M. Value-based sleep and breathing: Health economic aspects of obstructive sleep apnea. *Fac. Rev.* **2021**, *10*, 40. [[CrossRef](#)] [[PubMed](#)]
6. Richard, W.; Kox, D.; den Herder, C.; Laman, M.; van Tinteren, H.; de Vries, N. The role of sleep position in obstructive sleep apnea syndrome. *Eur. Arch. Oto-Rhino-Laryngol. Head Neck* **2006**, *263*, 946–950. [[CrossRef](#)] [[PubMed](#)]
7. Menon, A.; Kumar, M. Influence of body position on severity of obstructive sleep apnea: A systematic review. *Int. Sch. Res. Not.* **2013**, *2013*, 670381. [[CrossRef](#)]
8. Isono, S.; Shimada, A.; Utsugi, M.; Konno, A.; Nishino, T. Comparison of static mechanical properties of the passive pharynx between normal children and children with sleep-disordered breathing. *Am. J. Respir. Crit. Care Med.* **1998**, *157*, 1204–1212. [[CrossRef](#)] [[PubMed](#)]
9. Fallmann, S.; Chen, L. Computational sleep behavior analysis: A survey. *IEEE Access* **2019**, *7*, 142421–142440. [[CrossRef](#)]
10. Li, X.; Gong, Y.; Jin, X.; Shang, P. Sleep posture recognition based on machine learning: A systematic review. *Pervasive Mob. Comput.* **2023**, *90*, 101752. [[CrossRef](#)]
11. Enayati, M.; Skubic, M.; Keller, J.M.; Popescu, M.; Farahani, N.Z. Sleep posture classification using bed sensor data and neural networks. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 461–465.
12. Han, P.; Li, L.; Zhang, H.; Guan, L.; Marques, C.; Savović, S.; Ortega, B.; Min, R.; Li, X. Low-cost plastic optical fiber sensor embedded in mattress for sleep performance monitoring. *Opt. Fiber Technol.* **2021**, *64*, 102541. [[CrossRef](#)]
13. Wong, D.W.-C.; Wang, Y.; Lin, J.; Tan, Q.; Chen, T.L.-W.; Zhang, M. Sleeping mattress determinants and evaluation: A biomechanical review and critique. *PeerJ* **2019**, *7*, e6364. [[CrossRef](#)]
14. Tam, A.Y.-C.; Zha, L.-W.; So, B.P.-H.; Lai, D.K.-H.; Mao, Y.-J.; Lim, H.-J.; Wong, D.W.-C.; Cheung, J.C.-W. Depth-Camera-Based Under-Blanket Sleep Posture Classification Using Anatomical Landmark-Guided Deep Learning Model. *Int. J. Environ. Res. Public Health* **2022**, *19*, 13491. [[CrossRef](#)]
15. Tam, A.Y.-C.; So, B.P.-H.; Chan, T.T.-C.; Cheung, A.K.-Y.; Wong, D.W.-C.; Cheung, J.C.-W. A Blanket Accommodative Sleep Posture Classification System Using an Infrared Depth Camera: A Deep Learning Approach with Synthetic Augmentation of Blanket Conditions. *Sensors* **2021**, *21*, 5553. [[CrossRef](#)] [[PubMed](#)]
16. Masek, M.; Lam, C.P.; Tranthim-Fryer, C.; Jansen, B.; Baptist, K. Sleep monitor: A tool for monitoring and categorical scoring of lying position using 3D camera data. *SoftwareX* **2018**, *7*, 341–346. [[CrossRef](#)]
17. Ren, W.; Ma, O.; Ji, H.; Liu, X. Human posture recognition using a hybrid of fuzzy logic and machine learning approaches. *IEEE Access* **2020**, *8*, 135628–135639. [[CrossRef](#)]
18. Cheung, J.C.-W.; So, B.P.-H.; Ho, K.H.M.; Wong, D.W.-C.; Lam, A.H.-F.; Cheung, D.S.K. Wrist accelerometry for monitoring dementia agitation behaviour in clinical settings: A scoping review. *Front. Psychiatry* **2022**, *13*, 913213. [[CrossRef](#)] [[PubMed](#)]
19. Eyobu, O.S.; Kim, Y.W.; Cha, D.; Han, D.S. A real-time sleeping position recognition system using IMU sensor motion data. In Proceedings of the 2018 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 12–14 January 2018; pp. 1–2.
20. Davoodnia, V.; Etemad, A. Identity and posture recognition in smart beds with deep multitask learning. In Proceedings of the 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 6–9 October 2019; pp. 3054–3059.
21. Demiris, G.; Hensel, B.K.; Skubic, M.; Rantz, M. Senior residents’ perceived need of and preferences for “smart home” sensor technologies. *Int. J. Technol. Assess. Health Care* **2008**, *24*, 120–124. [[CrossRef](#)] [[PubMed](#)]
22. Otero, M. Application of a continuous wave radar for human gait recognition. In Proceedings of the Signal Processing, Sensor Fusion, and Target Recognition XIV, Orlando, FL, USA, 25 May 2005; pp. 538–548.
23. Kebe, M.; Gadhafi, R.; Mohammad, B.; Sanduleanu, M.; Saleh, H.; Al-Qutayri, M. Human vital signs detection methods and potential using radars: A review. *Sensors* **2020**, *20*, 1454. [[CrossRef](#)]

24. Lee, Y.; Park, J.-Y.; Choi, Y.-W.; Park, H.-K.; Cho, S.-H.; Cho, S.H.; Lim, Y.-H. A novel non-contact heart rate monitor using impulse-radio ultra-wideband (IR-UWB) radar technology. *Sci. Rep.* **2018**, *8*, 13053. [[CrossRef](#)]
25. Yim, D.; Lee, W.H.; Kim, J.I.; Kim, K.; Ahn, D.H.; Lim, Y.-H.; Cho, S.H.; Park, H.-K.; Cho, S.H. Quantified activity measurement for medical use in movement disorders through IR-UWB radar sensor. *Sensors* **2019**, *19*, 688. [[CrossRef](#)] [[PubMed](#)]
26. Ahmed, S.; Cho, S.H. Hand Gesture Recognition Using an IR-UWB Radar with an Inception Module-Based Classifier. *Sensors* **2020**, *20*, 564. [[CrossRef](#)] [[PubMed](#)]
27. Rana, S.P.; Dey, M.; Ghavami, M.; Dudley, S. Markerless gait classification employing 3D IR-UWB physiological motion sensing. *IEEE Sens. J.* **2022**, *22*, 6931–6941. [[CrossRef](#)]
28. Lai, D.K.-H.; Zha, L.-W.; Leung, T.Y.-N.; Tam, A.Y.-C.; So, B.P.-H.; Lim, H.-J.; Cheung, D.S.K.; Wong, D.W.-C.; Cheung, J.C.-W. Dual ultra-wideband (UWB) radar-based sleep posture recognition system: Towards ubiquitous sleep monitoring. *Eng. Regen.* **2023**, *4*, 36–43. [[CrossRef](#)]
29. Piriyaikitakonkij, M.; Warin, P.; Lakhan, P.; Leelaarporn, P.; Pianpanit, T.; Kumchaiseemak, N.; Suwajanakorn, S.; Niparnan, N.; Mukhopadhyay, S.C.; Wilaiprasitporn, T. SleepPoseNet: Multi-View Learning for Sleep Postural Transition Recognition Using UWB. *arXiv* **2020**, arXiv:2005.02176. [[CrossRef](#)]
30. Kiriazi, J.E.; Islam, S.M.M.; Borić-Lubecke, O.; Lubecke, V.M. Sleep Posture Recognition With a Dual-Frequency Cardiopulmonary Doppler Radar. *IEEE Access* **2021**, *9*, 36181–36194. [[CrossRef](#)]
31. Zhou, T.; Xia, Z.; Wang, X.; Xu, F. Human Sleep Posture Recognition Based on Millimeter-Wave Radar. In Proceedings of the 2021 Signal Processing Symposium (SPSymo), Łódź, Poland, 20–23 September 2021; pp. 316–321.
32. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Online, 13–19 June 2020; pp. 11534–11542.
33. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [[CrossRef](#)]
34. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
35. Islam, K. Recent advances in vision transformer: A survey and outlook of recent work. *arXiv* **2022**, arXiv:2203.01536.
36. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. *arXiv* **2020**, arXiv:2005.12872. [[CrossRef](#)]
37. Wang, H.; Zhu, Y.; Adam, H.; Yuille, A.; Chen, L.-C. MaX-DeepLab: End-to-End Panoptic Segmentation with Mask Transformers. *arXiv* **2020**, arXiv:2012.00759. [[CrossRef](#)]
38. Chen, Y.-S.; Cheng, K.-H.; Xu, Y.-A.; Juang, T.-Y. Multi-Feature Transformer-Based Learning for Continuous Human Motion Recognition with High Similarity Using mmWave FMCW Radar. *Sensors* **2022**, *22*, 8409. [[CrossRef](#)]
39. Huang, L.; Tan, J.; Liu, J.; Yuan, J. Hand-transformer: Non-autoregressive structured modeling for 3d hand pose estimation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXV 16. pp. 17–33.
40. Zhou, Y.; Xu, C.; Zhao, L.; Zhu, A.; Hu, F.; Li, Y. CSI-Former: Pay More Attention to Pose Estimation with WiFi. *Entropy* **2023**, *25*, 20. [[CrossRef](#)] [[PubMed](#)]
41. Chen, S.; He, W.; Ren, J.; Jiang, X. Attention-Based Dual-Stream Vision Transformer for Radar Gait Recognition. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 3668–3672.
42. Mogan, J.N.; Lee, C.P.; Lim, K.M.; Muthu, K.S. Gait-ViT: Gait Recognition with Vision Transformer. *Sensors* **2022**, *22*, 7362. [[CrossRef](#)]
43. Rahali, A.; Akhloofi, M.A. End-to-End Transformer-Based Models in Textual-Based NLP. *AI* **2023**, *4*, 54–110. [[CrossRef](#)]
44. Li, H.; Huang, J.; Ji, S. Bearing fault diagnosis with a feature fusion method based on an ensemble convolutional neural network and deep neural network. *Sensors* **2019**, *19*, 2034. [[CrossRef](#)]
45. Cuenat, S.; Couturier, R. Convolutional neural network (cnn) vs vision transformer (vit) for digital holography. In Proceedings of the 2022 2nd International Conference on Computer, Control and Robotics (ICCCR), Shanghai, China, 18–20 March 2022; pp. 235–240.
46. Kyathanahally, S.P.; Hardeman, T.; Reyes, M.; Merz, E.; Bulas, T.; Brun, P.; Pomati, F.; Baity-Jesi, M. Ensembles of data-efficient vision transformers as a new paradigm for automated classification in ecology. *Sci. Rep.* **2022**, *12*, 18590. [[CrossRef](#)] [[PubMed](#)]
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
48. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv* **2016**, arXiv:1608.06993.
49. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* **2019**, arXiv:1905.11946.
50. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv* **2021**, arXiv:2103.14030.
51. Taewoong Um, T.; Pfister, F.M.J.; Pichler, D.; Endo, S.; Lang, M.; Hirche, S.; Fietzek, U.; Kulić, D. Data Augmentation of Wearable Sensor Data for Parkinson’s Disease Monitoring using Convolutional Neural Networks. *arXiv* **2017**, arXiv:1706.00527.
52. Iwana, B.K.; Uchida, S. An empirical survey of data augmentation for time series classification with neural networks. *PLoS ONE* **2021**, *16*, e0254841. [[CrossRef](#)]

53. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Kai, L.; Li, F.-F. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
54. Hyun, E.; Jin, Y.S.; Lee, J.H. Moving and stationary target detection scheme using coherent integration and subtraction for automotive FMCW radar systems. In Proceedings of the 2017 IEEE Radar Conference (RadarConf), Seattle, WA, USA, 8–12 May 2017; pp. 476–481.
55. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
56. Tan, M.; Le, Q.V. EfficientNetV2: Smaller Models and Faster Training. *arXiv* **2021**, arXiv:2104.00298.
57. Wang, W.; Li, X.; Yang, J.; Lu, T. Mixed Link Networks. *arXiv* **2018**, arXiv:1802.01808.
58. Liang, X.; Zhang, H.; Fang, G.; Ye, S.; Gulliver, T.A. An Improved Algorithm for Through-Wall Target Detection Using Ultra-Wideband Impulse Radar. *IEEE Access* **2017**, *5*, 22101–22118. [[CrossRef](#)]
59. Lee, S.H.; Lee, S.; Song, B.C. Vision Transformer for Small-Size Datasets. *arXiv* **2021**, arXiv:2112.13492. [[CrossRef](#)]
60. Mohammadi, S.M.; Alnowami, M.; Khan, S.; Dijk, D.J.; Hilton, A.; Wells, K. Sleep Posture Classification using a Convolutional Neural Network. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 1–4.
61. Qiu, L.; Huang, Z.; Wirström, N.; Voigt, T. 3DinSAR: Object 3D localization for indoor RFID applications. In Proceedings of the 2016 IEEE International Conference on RFID (RFID), Orlando, FL, USA, 3–5 May 2016; pp. 1–8.
62. Soekhoe, D.; van der Putten, P.; Plaat, A. On the Impact of Data Set Size in Transfer Learning Using Deep Neural Networks. In *Advances in Intelligent Data Analysis XV. IDA 2016. Lecture Notes in Computer Science*, Boström, H., Knobbe, A., Soares, C., Papapetrou, P., Eds.; Advances in Intelligent Data Analysis XV; Springer International Publishing: Cham, Switzerland, 2016; Volume 9897, pp. 50–60.
63. Cheung, J.C.-W.; Tam, E.W.-C.; Mak, A.H.-Y.; Chan, T.T.-C.; Zheng, Y.-P. A night-time monitoring system (eNightLog) to prevent elderly wandering in hostels: A three-month field study. *Int. J. Environ. Res. Public Health* **2022**, *19*, 2103. [[CrossRef](#)]
64. Cheung, J.C.; Tam, E.W.; Mak, A.H.; Chan, T.T.; Lai, W.P.; Zheng, Y.P. Night-Time Monitoring System (eNightLog) for Elderly Wandering Behavior. *Sensors* **2021**, *21*, 704. [[CrossRef](#)]
65. Alp Güler, R.; Neverova, N.; Kokkinos, I. DensePose: Dense Human Pose Estimation In The Wild. *arXiv* **2018**, arXiv:1802.00434.
66. Geng, J.; Huang, D.; De la Torre, F. DensePose From WiFi. *arXiv* **2022**, arXiv:2301.00250.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.