SaccpaNet: A Separable Atrous Convolution-based Cascade Pyramid Attention Network to Estimate Body Landmarks Using Cross-modal Knowledge Transfer for Under-blanket Sleep Posture Classification

Andy Yiu-Chau Tam, Ye-Jiao Mao, Derek Ka-Hei Lai, Andy Chi-Ho Chan, Daphne Sze Ki Cheung, William Kearns, Duo Wai-Chi Wong, *Member, IEEE*, and James Chung-Wai Cheung, *Member*, *IEEE*

***Abstract*—The accuracy of sleep posture assessment in standard polysomnography might be compromised by the unfamiliar sleep lab environment. In this work, we aim to develop a depth camera-based sleep posture monitoring and classification system for home or community usage and tailor a deep learning model that can account for blanket interference. Our model included a joint coordinate estimation network (JCE) and sleep posture classification network (SPC). SaccpaNet (Separable Atrous Convolution-based Cascade Pyramid Attention Network) was developed using a combination of pyramidal structure of residual separable atrous convolution unit to reduce computational cost and enlarge receptive field. The Saccpa attention unit served as the core of JCE and SPC, while different backbones for SPC were also evaluated. The model was cross-modally pretrained by RGB images from the COCO whole body dataset and then trained/tested using dept image data collected from 150 participants performing seven sleep postures across four blanket conditions. Besides, we applied a data augmentation technique that used intra-class mix-up to synthesize blanket conditions; and an overlaid flip-cut to synthesize partially covered blanket conditions for a robustness that we referred to as the Post-hoc Data Augmentation Robustness Test (PhD-ART). Our model achieved an average precision of estimated joint coordinate (in terms of PCK@0.1) of 0.652 and demonstrated adequate robustness. The overall classification accuracy of sleep postures (F1-score) was 0.885 and 0.940, for 7- and 6-class classification, respectively. Our system was resistant to the interference of blanket, with a spread difference of 2.5%.**

***Index Terms*— computer vision; deep learning; human activity recognition; image classification; sleep.**

I. INTRODUCTION

Sleep is an essential part of our daily lives and has a significant impact on our health and quality of life. Sleep deprivations and disorders have been associated with the development of diabetes, cardiovascular diseases, obesity, and depression, whereas sleep apnea may induce hypertension, stroke, and coronary heart disease [1]. Sleep studies are non-invasive diagnostic tools for sleep disorders and help identify the underlying causes of sleep deprivation. Sleep studies are typically conducted in specialized sleep laboratories, where participants are equipped with various instruments, often referred to as PSG (polysomnography), to measure brain activity, eye movement, heart rate, breathing pattern, etc. During sleep tests, participants are required to rest comfortably in a controlled sleep laboratory, albeit with perhaps more than a dozen sensors physically attached to their bodies to facilitate comprehensive data acquisition. The collected data requires expertise of trained professionals to analyze.

In addition to physiological measurements in PSG, sleep movement/posture/behaviour is one of the important

Manuscript received dd September 2023; revised dd mmm yyyy; accepted dd mmm yyyy. Date of publication dd mmm yyyy; date of current version dd mmm yyyy. This work was supported in part by General Research Fund (GRF) from the University Grants Committee of Hong Kong under Grants PolyU15223822, and in part by the Research Institute for Smart Ageing of The Hong Kong Polytechnic University under Grants P0039001. Andy Yiu-Chau Tam and Ye-Jiao Mao are co-first authors. Duo Wai-Chi Wong and James Chung-Wai Cheung contributed equally to this work. (*Corresponding author: Duo Wai-Chi Wong, James Chung-Wai Cheung*)

Andy Yiu-Chau Tam and James Chung-Wai Cheung are with the Department of Biomedical Engineering, Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China, and Research Institute of Smart Ageing, The Hong Kong Polytechnic University, Hong Kong 999077, China (andy-yiu-chau.tam@connect.polyu.hk; james.chungwai.cheung@polyu.edu.hk).

Ye-Jiao Mao, Derek Ka-Hei Lai, Andy Chi-Ho Chan, and Duo Wai-Chi Wong are with the Department of Biomedical Engineering, Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China (yejiao.mao@connect.polyu.hk; derekkh.lai@connect.polyu.hk; andy-chi-ho.chan@connect.polyu.hk; duo.wong@polyu.edu.hk).

Daphne Sze Ki Cheung is with the School of Nursing, The Hong Kong Polytechnic University, Hong Kong 999077, China, and Research Institute of Smart Ageing, The Hong Kong Polytechnic University, Hong Kong 999077, China (daphne.cheung@polyu.edu.hk).

William Kearns is with the College of Behavioral and Community Sciences, Department of Child and Family Studies, University of South Florida, Tampa, FL 33612, USA (kearns@usf.edu).

The source code and depth image dataset are publicly available at https://github.com/bmeailabhkpu/SaccpaNet [available upon publish].

Digital Object Identifier [DOI_____]

assessments in sleep test. Common sleep-related movement disorders include sleep bruxism and restless leg syndrome, while individuals with parasomnia might exhibit aberrant sleep behaviours, such as sleepwalking, sleep-talking, nightmares, partial awakening and paralysis. One of the most prevalent sleep problems is insomnia; more than one-fifth of the general population may suffer from insomnia [2]. Insomnia is associated with poor sleep quality and poor sleeping posture. A good sleep posture could maintain the spine at a physiological curvature and enable muscle and soft tissue to relax and recover [3]. In contrast poor sleeping position/posture may lead to sleep-related musculoskeletal disorders or exacerbate existing neck, back, and joint pain [3]. Monitoring sleep posture and behaviour can also be used for non-medical purposes. For instance, it could be used for evaluating the ergonomics of sleep environments and mattresses [3], providing boundary conditions to simulate sleep and sleep disorders [4], as well as for surveillance of nighttime wandering behaviour of older adults [5, 6].

Nevertheless, there are some constraints in the sleep posture measurements in PSG system. The sleep test in a sleep laboratory itself could impact the sleep "performance" [7]. Sleeping in a new environment may create the "first night effect" that diminishes deep sleep and makes it difficult to fall asleep, and unfamiliar beddings (mattress and pillow) might cause people to adopt unusual sleep postures, positions and result in more turns and restlessness [7]. Traditionally, the sleep postures and behaviours are taped by video cameras overnight and manually examined. The process is labour-intensive, requires preparation time, and may violate privacy considerations. Intervening blankets or quilts in the dark environment can also affect the accuracy of the analysis. Some studies switched to the use of wearable devices with accelerometers (i.e., actigraphy) [8, 9]. However, they cause uncomfortable sleep, affect sleep quality, and introduce compliance problems, especially for older adults and those with related behavioural issues [10, 11].

We aim to develop a depth-camera based system with deep learning models to classify sleep postures as our first milestone. The novelty of this study lies in the development of a deep learning model, named as SaccpaNet (Separatable Atrous Convolution-based Cascade Pyramid Attention Network), that could classify sleep postures under a blanket. The core of SaccpaNet is a module for identifying musculoskeletal joints and segments based on a cascading pyramid attention unit for sleep posture classification. The unit could also be applied as the backbone in the subsequent classification stage.

In summary, the main contribution of this study is the development of SaccpaNet, which utilizes the concept of receptive field to enhance the accuracy of sleep posture classification. This is a significant advancement, as existing studies often overlook the influence of blankets or only consider a single blanket thickness. It is important to note that SaccpaNet employs cross-modal training on RGB data but uses depth images as input, addressing the issue of lights being turned off during sleep. Furthermore, we previously proposed a blanket synthetic technique for data augmentation to enhance model generalizability. In this study, we introduce a post-hoc data augmentation robustness test (Ph-DART) to assess model robustness using a dataset augmented by participants partially covered by blankets.

II. Related Work

Non-contact optical sensors, especially depth cameras, have overtaken sensor-mattresses as the preferred solution. However, a new issue has arisen with these sensors, as they can be affected by interference from blankets or quilts, and its impact is frequently assessed. Using a Kinetc Artec scanner, Ren, et al. [12] achieved an accuracy of 92.5% in classifying six sleep postures using Support Vector Machine (SVM) on scale-invariant feature transform (SIFT) features. Convolutional Neural Networks (CNN) were among the most widely used deep learning models for posture classification using depth cameras, but a drop of accuracy was also observed from sleepers under blankets [13, 14]. Besides, Li, et al. [10] applied the CNN on streamlined data from Kinetic depth camera and RGB camera to classify 10 sleep postures and found that the accuracy loss due to blanket covering was reduced to 3%. To accommodate the blanket issues, Tam, et al. [15] estimated the anatomical landmarks using an open-source pose estimator model from non-blanket conditions and superimposed these landmark coordinates on the data with blanket for model training. Their techniques improved the classification accuracy of ECA-Net50 from 87.4% to 92.2%, with less influence by blanket conditions.

Some other machine/deep learning models have been designed for sleep posture estimations. Yue, et al. [16] utilized a multilayer fully connected neural network to analyze a snapshot of multipath features of radiofrequency signals for predicting sleep postures. Building on the use of sleep video data, Li, et al. [17] introduced the SleePose-FRCNN-Net, a deep multitask learning network that combines a ResNet feature extractor with a Region Proposal Network. The network is triggered by a motion detector, identifying a bounding box to classify upper-body and head poses. Another study employed tensor factorization for dimensional reduction on infrared images and coupled with a pre-trained VGG19 network to classify under-blanket sleep postures [18]. Liu, et al. [19] proposed an innovative infrared selective image acquisition technique to mitigate the effect of lighting variation. In addition, they processed the images with a 2-end histogram of oriented gradient rectification to extract features accurately from participant locations. To estimate postures, they fine-tuned a multi-stage CNN structure using three strategies: MANNE-S6, MANNE-AS, and MANNE-AS-S2C3, each targeting different layers and stages of the network for optimization. Nevertheless, while existing studies have tested blanket conditions, they have not been dedicated to handle the complexity of blanket problems in deep.

III. Model Architecture

A. Overview

The overall model framework consists of two parts: the joint coordinate estimation (JCE) network for feature extraction and the sleep posture classification (SPC) network for classification (Fig. 1)

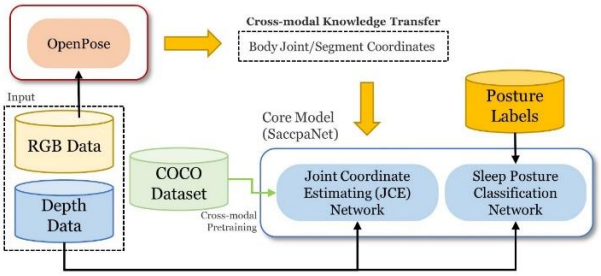

Fig. 1. Overview of the multimodal transfer learning in the overall framework.

The JCE network is based on an optimized encoder-decoder network to predict the positions of body joint/segment coordinates through depth images. The SPC network is laid by the convolutional residual network to classify sleep postures based on the depth images and the coordinate information estimated by the JCE network.

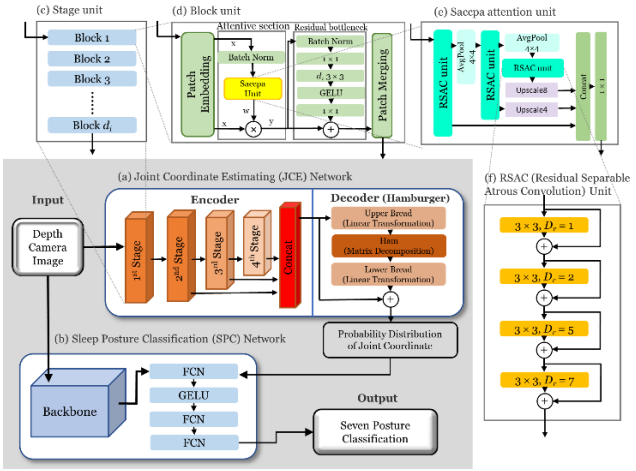

Fig. 2. Overview of Model Architecture of SaccpaNet: (a) Joint Coordinate Estimating (JCE) network; (b) Sleep Posture Classification (SPC) network; (c) stage unit under JCE; (d) block unit under stage unit; (e) Saccpa attention unit under block unit; (f) RSAC (residual separable atrous convolution) unit under Saccpa attention unit.

We apply a cross-modal knowledge transfer scheme on JCE (Fig. 1). The JCE is (cross-modally) pretrained with the RGB data and labelled coordinates of the COCO dataset [20]. Next, the depth images and the RGB-data-driven knowledge (i.e., coordinates) of our experiment is paired and superimposed to train JCE. The RGB-data-driven knowledge is generated by an existing network (OpenPose) [21] that was fed RGB data of our experiment to estimate the body joint/segment coordinates. In other words, RGB-D data are required for model pretraining and training but only depth images are needed for deployment.

B. Joint Coordinate Estimation (JCE) Network

The JCE network takes reference to SegNeXt [22] but has replaced its channel attention to Saccpa unit (Fig. 2a). The encoder of JCE enables feature extraction with a pyramid structure of 4 stages (Fig 2b). Each stage progressively reduces the resolution by half through passing a sequence of blocks of equal number of channels (Fig. 2c). Each block consists of a patch embedding unit, an attention unit (Saccpa unit), and operations resembling the residual bottleneck layer (Fig. 2d). The data patches are encoded with overlapping bits. Each data patch of $N \times N$ pixels are transformed into a vector of size $1 \times N^2$. The patch embedding is controlled by the kernel size (K), stride between two adjacent patches (S), and padding size (P). We assigned K = 3, S = 2, and P = 1 for the 1st stage; K = 8, S = 4, P = 3 for the 2nd stage [22].

As shown in Fig. 2d, the embedded patch vectors subsequently undergo residual batch normalization attention, which involves batch normalization and attention (through Saccpa unit). Next, the residual batch normalization convolution is performed, which involves batch normalization, a 1×1 input channel convolution, a 3×3 depth separated convolution, a GELU activation function, and finally a 1×1 output channel convolution layer. This arrangement of residual batch normalization convolution configuration resembles the bottleneck layer of ResNet [23]. Lastly, the patch merging process reverse-transforms the data to patches of $N \times N$ pixels.

The Saccpa unit contains a pyramidal structure of residual separable atrous convolution (RSAC) units, average pooling, and bilinear upscaling (Fig. 2e), which seek to extend the receptive field for generating long range relationships while maintaining resolution size. The output is then concatenated, followed by a 1×1 convolution for channel reduction.

There are four convolution processes in the RSAC units in the Saccpa unit (Fig. 2f). Fig. 3 explains the structure and operation of the convolution process. Both data input and kernels could be viewed as a structure with heights, width, and channel ($H \times W \times M$). Depthwise (with spatial separation) convolutions on height and width kernels, followed by a pointwise convolution, are implemented on the kernel, illustrated in equation (3) [24]. Atrous convolutions are poised on the spatial and depthwise convolutions at different dilation rates (D_r = 1, 2, 5, 7) and are summed (Fig. 2f). The computed attention weights of the Saccpa unit are element-wise multiplied with the input, equation (1) and equation (2).

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3432195

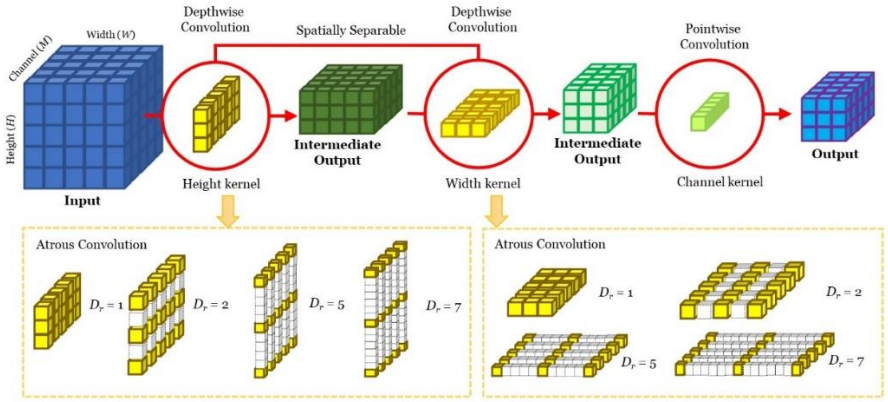

Fig. 3. Schematic diagram illustrating the convolution processes within the Residual Separable Atrous Convolution (RSAC) unit.

$$w = SACCPA(BatchNorm(x)) \tag{1}$$

$$y_{ij} = x_{ij} * w_{ij} \tag{2}$$

$$O_{i.j.n} = \sum_{m=1}^{n} \left(\sum_{d=1-p}^{i+p} \sum_{e=1-p}^{j+p} M_{d,e,m} K_{d,1,m} K_{1,e,m} \right) P_m \tag{3}$$

Here, x denotes the embedded patch vectors to the attention module, y denotes the attended output, w denotes the attention weights which indicates the importance of each feature, x_{ij} , y_{ij} , w_{ij} denotes the value of x , y , w respectively. O is the output feature of the separable atrous convolution with spatial indices height (i), width (j), and channel (n). M is the input feature with spatial indices height (d), width (e), and channel (m). K is the kernel while P is the pointwise convolution with padding p.

The attended output is passed to the residual bottleneck layers, followed by patch merging. The patch embedding, attention unit, residual bottleneck layers and patch merging forms a block, and blocks are cascaded to form stages.

At the end, after the iteration of the process in the four stages, the 2nd, 3rd and 4th stages of the encoder is concatenated and fed to the decoder (Fig. 2a). The 1st layer is excluded, as too much low-level features may reduce model performance. The Hamburger decoder [25] is adopted, which consists of upper and lower bread layers of linear transformation sandwiching a ham layer of matrix decomposition. It could factorize the learnt representation into sub-matrices in order to retrieve the low-rank signal subspace with clean data [25]. The drop rate of the Hamburger decoder is set to 0.1. the output of the decoder is arranged in form of a two-dimensional heatmap that represents the probability distribution of each body joint/segment coordinates on each pixel. The predicted coordinates for each body joint/segment point (coordinate) are established by the Soft-Argmax operation on the heatmap.

C. Sleep Posture Classification (SPC) Network

The SPC consists of a backbone, followed by a classification head, as shown in Fig. 2b. In this study, we evaluated different kinds of networks (and number of layers) as backbone, including ResNet (layer = 34, 50, 101, 152), ECANet (layer = 34, 50, 101, 152), EfficientNet (scaling = B0, B2, B4, B7), and ViT (model/patch size = B/16, B/32, L/16, L/32). In addition, we introduced another backbone that employs channel attention with the Saccpa unit (Fig. 2e) in the convolution network (layer = 34, 50, 101, 152). We dubbed it as CNN-Saccpa to better distinguish it from the overall network (SaccpaNet).

The outputs from the backbone are incorporated with the predicted segment/joint coordinates from JCE in the FCN layer of the classification head, followed by a GELU activation, and two FCN layers to finalize the seven-posture classification (Fig. 2b). Besides, we have utilized ResNet152-SVM (without JCE) as the baseline for comparison. ResNet152 is pretrained using ImageNet1K and trained to produce a feature of 2048 dimensions from depth images. Subsequently, SVM is employed as a classifier.

IV. Knowledge Transfer, Hyperparameter Tuning, and Pretraining

A. Cross-modal Knowledge Transfer

Cross-modal knowledge transfer is facilitated by OpenPose, which is a real-time multi-person human pose detection library [21]. It uses a part affinity field to estimate the degree of association of human parts [21]. It can estimate (output) coordinates of 18 anatomical landmarks, including hand, foot, facial, etc., from RGB image data [21]. The annotation of the body joint/segment coordinates is then superimposed and transferred to the corresponding depth image data with and without blankets, since the participants maintained the same posture under different blanket conditions. The depth image data with labelled coordinates are used to train JCE.

B. Hyperparameter Tuning

Hyperparameter tuning of the model architecture has been performed on JCE based on an existing network search space design [26]. Its input are depth images, whilst the ground truth (coordinates) is transferred from RGB data through OpenPose (Fig. 1). The total number of blocks (i.e.,

depth) of a particular stage i, is denoted as d_i, and the total number of blocks for all stages is denoted as $d = \sum_{i=1}^{4} d_i$. The width of the j^{th} block ($0 \leq j < d$), u_j is presented as a linear parameterization of initial width, $w_0 > 0$ and the coefficient, $w_a > 0$ in the equation (4). u_j is then quantized into per-block width, w_j in equation (5) by computing a width multiplier, $w_m > 0$ and $\lfloor s_j \rceil$ for each block j.

$$u_j = w_0 + w_a * j = w_0 \cdot w_m^{s_j} \quad (4)$$

$$w_j = w_0 \cdot w_m^{\lfloor s_j \rceil} \quad (5)$$

To re-format the equation per-lock level, each stage i has a block width, $w_i = w_0 \cdot w_m^i$, and number of blocks, $d_i = \sum_j \mathbf{1}[\lfloor s_j \rceil = i]$, where "**1**" is an indicator function that takes one when the condition is met.

Since the output of JCE are the estimated coordinates of the body joint/segment positions, the evaluation is based on the accuracy measure, percentage of correct key-points transfer (PCK) with a tolerance factor θ, as shown in equation (6). It measures the percentage of predicted coordinates with an offset distances less than the tolerance factor, which is the percentage length of the diagonal of the bounding box of the full pose estimation region. The tolerance factor θ is often set to 0.1 [4].

$$PCK(\theta) = \frac{\sum_{k=1}^{n} \mathbf{1}\left[\left\| \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix}, \begin{pmatrix} x \\ y \end{pmatrix} \right\|_2 < \theta \cdot d_{BB}\right]}{\sum_{k=1}^{n} k} \quad (6)$$

where d_{BB} denotes diagonal distance of bounding box and the double stroke denotes Euclidean distance.

Random search has been conducted over the combinations of d, w_0, w_a, and w_m. In total, there are 4 to 14 million of hyperparameters in the sample space. Table I represents the space for the random search. The model was run for 300 epochs with a batch size of 16 using the Adam optimizer with a learning rate of $5e^{-5}$.

TABLE I
SEARCH SPACE AND RESULTS OF HYPERPARAMETER TUNING

Parameter	Range	Interval	Sampling Distribution	Tuned Results
d	12 – 28	1	Uniform	17
w_0	8 – 256	8	Log Uniform	8
w_a	8 – 256	0.1	Log Uniform	12.8
w_m	2 – 3	0.001	Log Uniform	2.942

From the results of random search, we found that the precision accuracy was unlikely to be associated with the change of d and w_0. Therefore, we further conducted a grid search on the other parameters, w_a and w_m at a higher precision, while keeping $d = 17$ and $w_0 = 8$. The former (w_a) was searched from 12.6 to 13.0 at an interval of 0.1, while the latter (w_m) was searched from 2.9 to 3.0 at an interval of 0.02. The grid search used the same number of epoch and batch size as that of the random search. The tuned hyperparameters were shown in Table I. It yielded a PCK@0.1 value of 0.6616 on the validation set.

C. Cross-modal Pretraining

A cross-model pretraining strategy is applied to pretrain our model (JCE) with RGB data and then trains it on depth image data based on the assumption that they have semantically comparable encoded representations. The pretraining is facilitated by the COCO-whole body dataset (COCO) [27]. COCO-whole body is a big dataset of 200k images with 250k person instances. It consists of in-the-wild RGB images of the whole human body labelled with 133 key body joint and segment landmarks.

From the 133 joints, we have extract 18 of them, which aligns with that of OpenPose (detailed in Section IV-C). We have converted the RGB images of the datasets to greyscale to discard the color information that is also unavailable for depth images. Images are downsized to 192 × 256 pixels.

Data augmentation is conducted to the dataset during the pretraining, which included horizontal random flips, random half body operations, random bounding box transformations (position, rotation, and scale), and affine transformations.

A separate decoder head is applied on the dataset using the MMPose's implementation of simple deconvolution head [28] that could produce a heatmap data of size 48 × 64 pixels with a Gaussian variance (σ) of 2 pixels. The pretraining is conducted using Pytorch 1.11.0 and the MMPose library. The process is facilitated by an Adam optimizer. A total of 210 epochs of batch size 32 is trained using an initial learning rate of $3.125e^{-5}$ for the first 170 epochs, a 10-fold decay for the following 30 epochs, and a further 10-fold decay for the final 10 epochs.

V. MODEL TRAINING AND EVALUATION

A. Experimental Protocol and Data Collection

We recruited 150 healthy participants with a fair distribution of age (mean: 40.6, SD: 21.0, range: 17 – 77) and gender (77 males and 73 females) for data collection. Their mean height and body weight were 165.6 cm (SD: 9.41 cm) and 60.9 kg (SD: 11.53 kg), respectively. Participants were excluded if they reported severe sleep deprivation, sleep disorder, pain, or musculoskeletal problems. The experiment and protocol were approved by the Institutional Review Board (reference number: HSEARS20210127007). All participants signed an informed consent after receiving oral and written descriptions of experimental procedures before the start of experiment.

RGB-D data were collected by an active infrared stereo technology depth camera (Realsense D435i, Intel Corp., Santa Clara, CA, United States) that incorporated with an auxiliary visible light RGB camera and an inertia measurement unit (IMU). The IMU was not utilized in this study. The resolution of the camera was 848 × 480 pixels with a sampling frequency of 6 fps and installed 1.6 m above a standard hospital bed of 55 cm tall.

The participants were instructed to lie in seven sleep (recumbent) postures in the following orders: (1) supine, (2)

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3432195

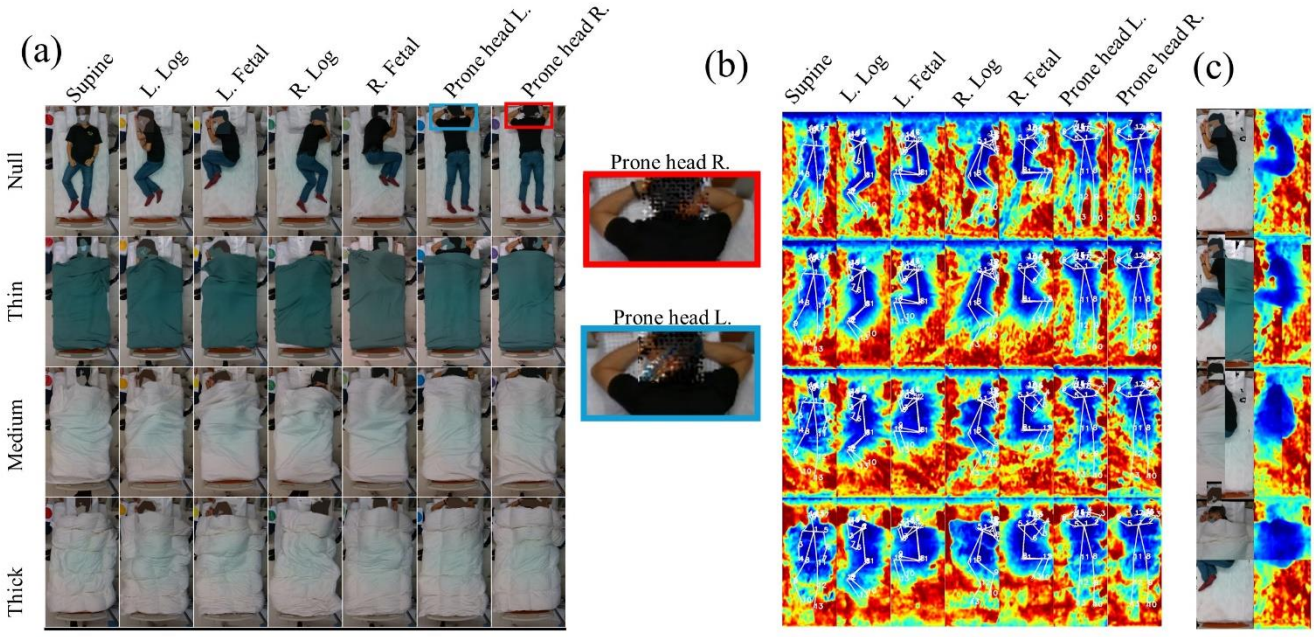

Fig. 5. Demonstration of experimental protocol of seven postures over four blanket conditions with: (a) optical camera; (b) depth camera overlaid with JCE results; (c) demonstration of overlaid flip cut to synthesize partial blanket coverings for data augmentation.

prone with head turned left, (3) prone with head turned right, (4) log left, (5) log right, (6) fetal left, and (7) fetal right (Fig. 5) and were manually labelled in the dataset. The prone postures were the same except the head direction (Fig. 5a). For each posture, we imposed four blanket conditions by covering blankets over the participants (except head) with blankets of varying thicknesses (thick, medium, thin, and none) in sequence. The participants remained stationary when our researchers covered them with different blankets. All blankets were sourced from IKEA (Delft, Netherlands), which were, respectively, *Fjallarnika* extra warm duvet (8 cm thick), *Saffero* light warm duvet (2 cm thick), and *Vallkrassing* duvet (0.4 cm thick). Participants were given time to adjust their self-selected most comfortable position for each posture before data collection. Sleep postures and blanket conditions were labelled manually.

B. Model Training

There were 28 paired sets (4 blanket conditions × 7 sleep postures) of streamed RGB-D data for each participant, resulting in a total of 4,200 RGB-D pairs of images. The model training-testing-validation ratio was 64:20:16. The validation set was applied for hyperparameter tuning, as described in the previous sections.

The cross-model pretrained learned weights by the COCO dataset were transferred to our model (JCE). Our training set depth data were used to fine-tune (train) the JCE model with transferred ground truth coordinates of RGB data (Fig. 1). The depth data and JCE model output of the training set were input to SPC for training and used the manually labelled posture as ground truth (Fig. 1). Only depth data of the testing set were utilized for prediction in model testing of the SaccpaNet but not the RGB data. Besides, to evaluate the performance of the JCE network alone, we retrieved the estimated joint coordinate data of JCE by the depth data of the testing set and compared them to the coordinates transferred from RGB testing set data through OpenPose.

Model training was implemented using the PyTorch Deep Learning Framework. The model training was conducted using Pytorch 2.0.1 with a batch size of 8 and a total of 600 epochs. Data augmentations with affine transformation, in addition to an intraclass mix-up algorithm for blanket synthesis [14] were proceeded at batch level during the model training. This process generated more blanket conditions (thicknesses) based on the available datasets of three blanket thicknesses to improve the generalizability of the model. Cross entropy loss was used as the objective function. Adam optimizer was set at a learning rate of 0.000025, L2 regularization of 0.00000025. Fig. 4 shows the loss curve of the training and testing dataset.

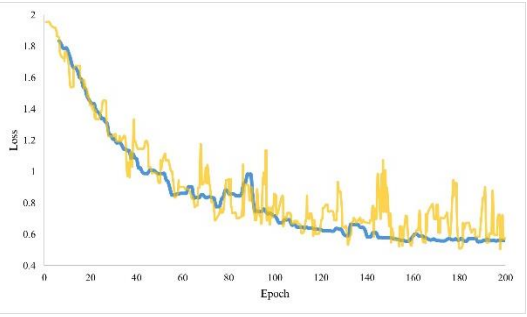

Fig. 4. An illustration of loss curve of SaccpaNet152 on the training (orange) and testing (blue) dataset.

C. Ablation Study

A unit-by-unit ablation study was conducted to evaluate the functions and effects of individual units in the Saccpa module. Model training and testing tasks were carried out by removing different levels of the model, including convolutions inside RSAC unit (Fig. 2f), RSAC units inside

Saccpa attention unit (Fig. 2e), and blocks inside stages (Fig. 2a and 2c). For ablation study on convolutions inside RSAC unit, we evaluated the removal of the last 1, 2, and 3, separable atrous convolution operations. For that on RSAC unit inside Saccpa attention unit, we evaluated the removal of last 1, 2, and 3 (all) RSAC units. Regarding the blocks within the stages, we evaluated six combinations, each with different number of blocks at different stages, while maintaining the pyramid structure. F1-scores were calculated for the ablation study using the validation subset.

D. Post-hoc Data Augmentation Robustness Test (PhD-ART)

In addition to using data augmentation to synthesize blanket conditions, we proposed and developed another augmented dataset to simulate scenarios where blankets were partially covering the body. Images of different blankets and without blankets were captured at the same position for each individual. Therefore, the blanket itself was extracted as a separate layer and can overlaid on the condition without a blanket. Subsequently, this blanket layer was then randomly cut in half, both vertically and horizontally, before being placed on the participants' images (Fig. 5c). We refer to this as the overlaid flip-cut technique.

The augmented dataset served as an external testing set to evaluate the robustness of the model that we named it as PhD-ART. For each posture and blanket condition, four pieces of augmented data were generated. F1-score and PCK@0.1 of the external testing was compared to that of model testing.

VI. Results

Table II presents the network sizes of various models with different backbones plus JCE, JCE alone, and the baseline model (ResNet152-SVM) without JCE. For the JCE model alone, the size of the multiply-accumulate operation (MAC) and the trainable parameters were 11.0B and 8.53M, respectively. The most lightweight network was the EfficientB0 model integrated with JCE, while the heaviest network was our proposed CNN-Saccpa152 model, also incorporating JCE. The table further illustrates the scalability of the CNN-Saccpa model. As the number of layers increases, the performance of the CNN-Saccpa model improves, a feature not observed in other models.

For JCE, the precision accuracy of joint coordinates on the testing set, in terms of PCK@0.1 value, was 0.6518. In other words, more than 65% of the (anatomical landmarks) joint coordinates were correctly predicted within 10% diagonal distance of the bounding box. The results of PCK are also illustrated and visualized in Fig. 5b.

Table III presents the F1-scores and one-vs-all AUC (area under the receiver-operating characteristic curve) for various postures and blanket conditions when different deep learning models were employed as the backbone of SPC. Fig. 6 shows the ROC curves, in which all of them demonstrated good to excellent discriminative power with AUC > 0.9 (Table II).

On the other hand, the baseline model's performance was substantially lower than that of the other models. Notable misclassifications were observed between the log and fetal postures, as well as among prone postures, which may indicate a failure to identify subtle changes in posture. A slight improvement was observed (F1 0.5435) when we consolidated the classes into left/right and prone postures. However, the performance was still not up to the mark.

Fine-grained analyses on sleep postures and blanket conditions were conducted on the best-performing model (i.e., backbone using CNN-Saccpa152), as shown in the confusion matrices in Fig. 7 and the radial barcharts in Fig. 8. The subgroup classification performance of SaccpaNet using CNN-Saccpa152 as backbone on sleep postures is shown in Table III, in addition to the F1-scores across different postures and blanket conditions.

TABLE II
MODEL SIZE AND PERFORMANCE OF MODELS WITH DIFFERENCE BACKBONE

Network	#MACs	#Params	F1	AUC
JCE only	11.0B	8.53M	-	-
*Baseline without JCE	-	-	0.4115	0.8560
EfficientNetB0+JCE	11.4B	14.1M	0.8561	0.9721
EfficientNetB2+JCE	11.7B	17.9M	0.8611	0.9752
EfficientNetB4+JCE	12.5B	28.1M	0.8602	0.9756
EfficientNetB7+JCE	16.2B	75.1M	0.8406	0.9737
ECANet34+JCE	14.6B	30.6M	0.8541	0.9734
ECANet50+JCE	15.1B	34.4M	0.8551	0.9722
ECANet101+JCE	18.7B	53.3M	0.8200	0.9637
ECANet152+JCE	22.4B	69.0M	0.8388	0.9655
ResNet34+JCE	14.6B	30.6M	0.8734	0.9760
ResNet50+JCE	15.1B	34.4M	0.8442	0.9728
ResNet101+JCE	18.7B	53.3M	0.8431	0.9722
ResNet152+JCE	22.4B	69.0M	0.8499	0.9755
ViT-B/16+JCE	22.3B	66.9M	0.8190	0.9683
ViT-B/32+JCE	14.0B	68.6M	0.8464	0.9801
ViT-L/16+JCE	50.9B	212M	0.8419	0.9782
ViT-L/32+JCE	21.2B	214M	0.8521	0.9824
CNN-Saccpa34+JCE	15.7B	37.3M	0.8221	0.9535
CNN-Saccpa50+JCE	31.3B	136M	0.8346	0.9590
CNN-Saccpa101+JCE	52.2B	246M	0.8718	0.9829
CNN-Saccpa152+JCE	**73.1B**	**337M**	**0.8849**	**0.9833**

AUC: Area under receiver-operating characteristic curve; MAC: multiply-accumulate operation; Params: trainable parameters.
*Baseline refers to ResNet152-SVM.

The accuracy (F1-score) of CNN-Saccpa152 in predicting supine, log, and fetal postures ranged from 0.897 to 0.970. However, the system network had a poor F1-score for prone postures (0.753 to 0.756), as shown in Table III. Therefore, we attempted to conduct another test that reduced the classification to prone position exclusively (i.e., 6 classes). In this case, the F1-score increased from 0.8849 (7-class average) to 0.9399 (6-class weighted average). The F1-score for predicting prone postures rose to 0.9301 (Table III).

Thick blankets influenced the accuracy of the posture prediction. The F1-scores were the lowest among all blanket conditions and were 0.8616 and 0.9282, respectively for the 7-class average and 6-class weighted average (Table III).

Notably, thin blanket conditions showed the best results and performed better than no blanket. The F1-scores were 0.9083 and 0.9534, respectively for the 7-class average and 6-class weighted average (Table III). Moreover, the network demonstrated resilience against blanket conditions. The spread difference of F1-score between blanket conditions was small (Fig. 7), 0.047 and 0.025, respectively for the 7-class average and 6-class weighted average.

TABLE III
PERFORMANCE OF SACCPA152 UNDER DIFFERENT POSTURES AND BLANKET CONDITIONS.

Posture class	F1-scores of different blanket conditions				Overall blanket conditions		
	Null	Thin	Medium	Thick	F1-score	Precision	Recall
7-class average*	0.8812	0.9083	0.8881	0.8616	0.8849	0.8881	0.8869
6-class weighted average	0.9388	0.9534	0.9387	0.9282	0.9399	0.9450	0.9393
Supine	0.9474	1.0000	0.9677	0.8889	0.9504	0.9426	0.9583
Right log	0.9091	0.8955	0.8788	0.9062	0.8973	0.8252	0.9833
Left log	0.9231	0.9524	0.9524	0.9677	0.9486	0.9023	1.0000
Right fetal	0.9286	0.9655	0.9655	0.9492	0.9524	0.9910	0.9167
Left fetal	0.9474	0.9655	0.9655	1.0000	0.9700	1.0000	0.9417
Prone (head left & right)	0.958	0.9474	0.9204	0.8929	0.9301	0.9771	0.8875
Prone (head left)	0.7541	0.7857	0.7451	0.6538	0.7364	0.8100	0.6750
Prone (head right)	0.7586	0.7931	0.7419	0.6667	0.7395	0.7458	0.7333

*Seven classes considered both prone postures with head left and right.

The ablation study revealed that the original model, when tested with the validation subset, exhibited the best performance (Table IV). Eliminating the separable atrous convolution operations and the Saccpa attention unit resulted in a decrease in the model's performance. Furthermore, the attenuation of the blocks within the stages could potentially have an even more significant impact, which decreased the F1-score from 0.7706 to a minimum of 0.7203.

Our robustness test via PhD-ART showed that PCK@0.1 was slightly decreased from 0.6518 to 0.6378 and the F1-score was also slightly decreased from 0.8879 to 0.8476. This demonstrated that the model was robust against adversarial conditions (i.e., conditions where blanket is partially covering).

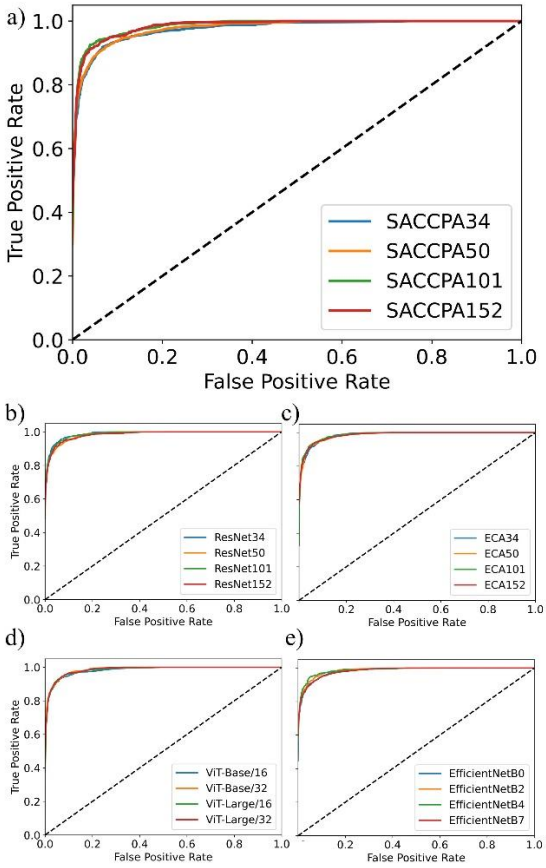

Fig. 6. Receiver-operating characteristics (ROC) curve of: (a) CNN-Saccpa; (b) ResNet; (c) ECANet; (d) ViT; and (e) EfficientNet.

TABLE IV
RESULTS OF ABLATION STUDY ON VALIDATION SUBSET

Ablation Level	Item Removed	F1
Original	N/A	**0.7706**
RSAC	Last SAC	0.7550
	Last 2 SAC	0.7502
	Last 3SAC	0.7560
Saccpa Attention	Last RSAC	0.7588
	Last 2 RSAC	0.7502
	Last 3 RSAC	0.7560
Stages at Blocks	1 block at stage 3	0.7501
	2 blocks at stage 3	0.7309
	2 blocks at stage 3 & 1 block at stage 4	0.7203
	2 blocks at stage 3 & 2 block at stage 4	0.7370
	2 blocks at stage 3 & 3 block at stage 4	0.7333

SAC: separable atrous convolution

VII. DISCUSSION

In this study, we proposed the SaccpaNet that aimed at enhancing context modelling by expanding the receptive field (i.e., input space) of attention units. We demonstrated that SaccpaNet can be employed as the core for joint coordinate estimation via cross-modal training and as the backbone for sleep posture classification. The key innovation of this network is to integrate different model architecture characteristics, including cascade pyramid network, separable convolution, and atrous convolution, as the attention module, which may ultimately broaden the receptive field at a relatively low computational cost. In addition, our work might represent one of the first datasets (n > 100) of depth camera images of different sleep postures.

A larger receptive field could enable the investigation of long-range spatial dependencies, which relate to the spatial relationship between two "points" far apart from each other in the input space. A previous study showed that the accuracy of posture estimation increased when long-range spatial dependencies and a larger receptive field were considered. Previous literature modelled contextual information based on global information by patching [29-31], pooling strategy [32-34], or increasing kernel size for convolutional attentions [22, 35]. Nevertheless, pooling strategies (e.g., global pooling) resulted in a loss of information, while the kernel size was usually pre-assigned and not adaptive to the input resolution. The attention mechanism is an adaptive selection process based on the input features [35], and our proposed attention mechanism was developed by the integration of the aforementioned architecture characteristics.

We combined channel-wise separable convolution and spatial sparable convolution to reduce computational cost [36-38]. Compared to conventional convolution, our approach could reduce the number of parameters from $C^2 K^2$ to $2C^2 K H$, and the number of flops from $C^2 K^2 H W$ to $(C^2 + 2K) H W$.

Atrous operation also enlarges the receptive field through inserting gaps between filter values to capture long-range dependencies without increasing the number of parameters [39]. With atrous operation, the size of the receptive field would grow exponentially as additional layers were stacked [40]. The gap distances of 1, 2, 5, and 7 were chosen because they were coprime numbers and helped reduce the gridding effect [41]. The resulting unit (RSAC) of our model covered a receptive field of 157 pixels, which is larger than that of the Large Kernel Attention (13 pixels) [35]. At last, the RSAC units were stacked in a multistage architecture design (a cascade pyramid network) [42] to incorporate contextual information under different scales.

With respect to the JCE performance, our hyperparameter tunning approach headed for a relatively lightweight model for joint coordinate estimation. The tuned model was subsequently pretrained by the COCO datasets. We achieved an overall PCK@0.1 of 0.652 after model training by our own dataset. Comparing to other recent models, ViTPose+-H has a PCK@0.1 of 0.759 [31], while that of a fine-tuned convolutional pose machine achieved a PCK@0.1 of approximately 0.62 (by observation of the result Figure) [19]. Considering the different in input modalities and blankets not fully accounted in other studies, we believed that our model performed well. The contribution of JCE on the classification performance was also illustrated in the unit-by-unit ablation study. The performance was the best when none of the units were removed. Table V presents a comparison of similar models employing Joint Coordinate Estimation (JCE). Utilizing the same classifier architecture for a fair comparison, our model exhibits a 4.8% reduction in performance relative to the one incorporating OpenPose for JCE. However, it is important to note that the latter requires both RGB and depth images for input, whereas our model operates solely on depth images. Moreover, our model's performance is on par with those not utilizing joint estimation. Significantly, our approach offers the advantage of providing supplementary information, such as limb placement, which can facilitate subsequent analysis. Variations in performance are expected due to differences in data inputs and model configurations.

TABLE V
COMPARISON OF MODEL PERFORMANCE (F1-SCORE) FOR SLEEP POSTURE CLASSIFICATION WITH JOINT ESTIMATION.

Study	Testing dataset	Joint Estimation	Classifier Backbone		
			EfficientNetB4	ResNet50	ECA-Net50
This study	D	JCE	0.860	0.844	0.885
[15]	RGB+D	OpenPose	0.908	0.913	0.922
[15]	D	No	0.873	0.836	0.874

D: depth images; RGB: red-green-blue images.

With regards to the posture classification performance, our best-performing model, Saccpa152 produced an overall F1-score of 0.8849. From the confusion matrix, we found that the model might not easily classify prone postures with head turning left and right. If we did not subclassify prone postures, the accuracy could be increased to 0.9301 F1-score, and the overall weighed-averaged F1-score could reach 0.938. On the other hand, we also demonstrated that our best-performing model was resilient against blanket conditions with the spread distance of F1-score of 2.5% in classifying 6 postures and 4.7% in classifying 7 postures. Interestingly, we also found that the classification performance for thin blanket was better than that of no blanket. The depth camera might not have the resolution to capture the abrupt depth differences between the boundaries of the body surface and the bed level. The thin blanket smoothed the spatial/depth discrepancies of the physical form and enhanced the data continuity. Besides, the improved performance under a thick blanket for particular postures (e.g., left log and fetal), as opposed to thinner blankets, may be attributed to the regularization-like effect. The thick blanket could serve as structural noise to prevent the model from overfitting the finer details of the body surface that are more variable and less discriminative for classification. The model then can focus on more primary patterns that are crucial for distinguishing postures.

In the field of machine learning, generalizability is the model's ability to maintain performance on new, unseen data, while robustness pertains to the model's stability against noisy or adversarial inputs. Data augmentation bolsters both robustness and generalizability. However, traditional methods using affine transformations may fall short in complex scenarios, such as varying blanket thicknesses. To this end, we have proposed two innovative data augmentation processes to this specific challenge. The first, an intra-class mix-up technique, was developed to synthesize a wider array of blanket data by blending images with different blanket thicknesses from the original dataset, thereby enhancing the model's generalizability to unfamiliar blanket conditions and reducing the risk of overfitting. The second, PhD-ART introduces the overlaid flip-cut technique. This approach generates scenarios where the subject is partially obscured by a blanket, thereby evaluating the model's resilience to such adversarial conditions. This serves as a robustness assessment for real-world applications.

The scalability of our model was demonstrated by a consistent enhancement in performance as additional layers are integrated. Our model might effectively utilize hierarchical features, which become increasingly abstract and informative with each added layer. Moreover, our model appears to be particularly well-adapted to the domain of under-blanket sleep posture classification, where nuanced features are more discernible at greater depths. In contrast, backbone models like ResNet experienced a decline in performance with added depth. This could be attributed to issues such as overfitting or the vanishing gradient problem, where the extra layers do not necessarily contribute to learning more useful representations.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3432195

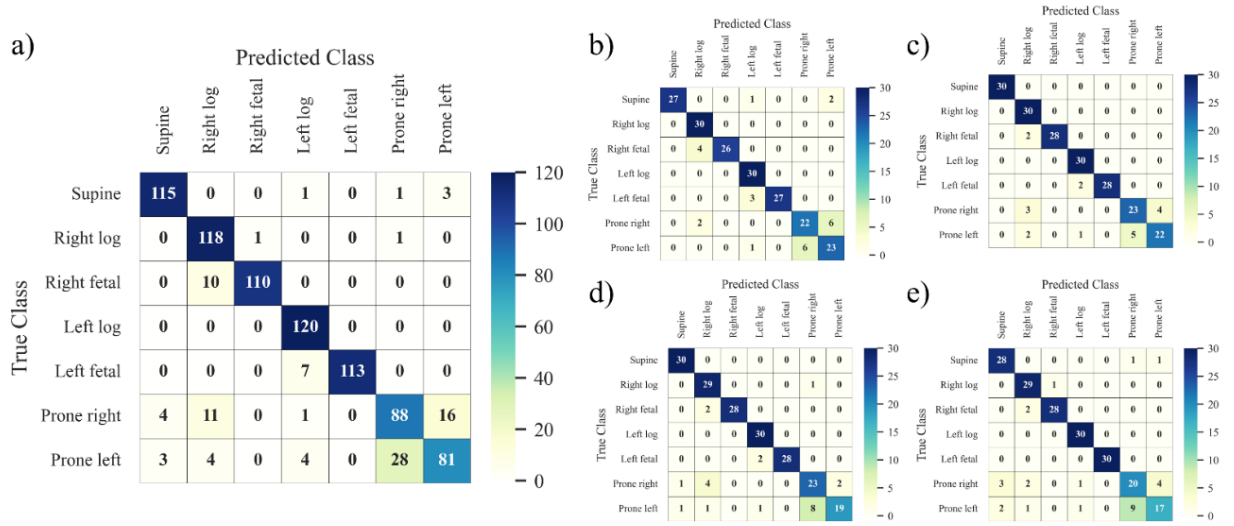

Fig. 7. Confusion matrices of sleep posture predictions of different blanket conditions: (a) overall; (b) no blanket;(c) thin; and (d) medium; (e) thick.

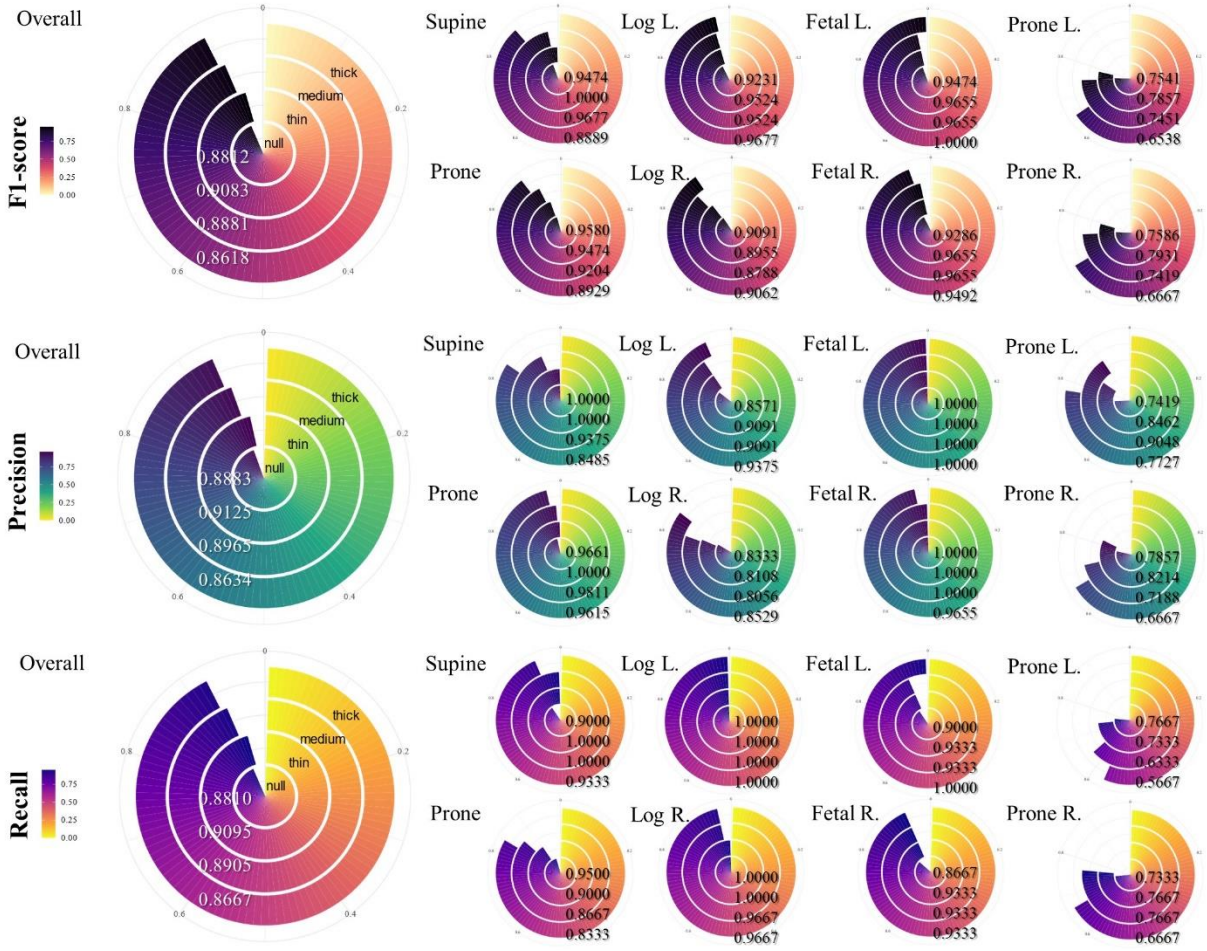

Fig. 8. Radial barcharts comparing classification performance between different blanket conditions overall and on each posture.

There were some limitations in this study. As stated previously, the classification accuracy of prone postures with head facing left and right was poor. This might be because prone postures hindered crucial facial features and the resolution of the depth camera may not be sensitive enough to account for the remaining features. Such a limitation could potentially be alleviated by incorporating handcrafted features, i.e. the kinematics of shoulders. Second, our model was restricted to detecting discrete events (i.e., static postures). A full deployable system shall be built in order to account for timeliness data and prediction (full sleeping duration) as well as posture changes (toss and turn), intermediate postures, and other on-bed behaviors and bed-exiting events [5, 6]. Future studies may consider background removal of the dataset and data augmentation using Generative Adversarial Network (GAN) or Generative

Artificial Intelligence (AI) to generate different mattress and blanket features to improve the robustness of the model. Besides, the attention mechanism can function as a tracking tool for other modalities, such as radar [43, 44] and microphone [45]. This enables these instruments to focus on their point of interest, such as apnea sounds, respiratory movements of the chest, and restless limb activities. This study lays foundation for comprehensive exploration of sleep quality studies and could potentially provide an alternative approach to conventional polysomnography.

VIII. Conclusion

We developed a novel model architecture, SaccpaNet, that included an attention network to expand receptive field at a relatively low computational cost. Even with the interference of blankets, SaccpaNet showed a rather high degree of precision in finding points of interest (PCK@0.1 = 0.6518). In addition, the classification accuracy of sleep postures was outstanding, with an overall F1-score of 0.8881 and 0.945, for 7-class and 6-class classification, respectively. Particularly, the overall F1-score reached 0.954 in thin blanket conditions.

References

[1] C. Arnaud, T. Bochaton, J.-L. Pépin, and E. Belaidi, "Obstructive sleep apnoea and cardiovascular consequences: pathophysiological mechanisms," *Archives of cardiovascular diseases*, vol. 113, no. 5, pp. 350-358, 2020.

[2] L.-N. Zeng *et al.*, "Gender difference in the prevalence of insomnia: a meta-analysis of observational studies," *Frontiers in Psychiatry*, vol. 11, p. 577429, 2020.

[3] D. W.-C. Wong, Y. Wang, J. Lin, Q. Tan, T. L.-W. Chen, and M. Zhang, "Sleeping mattress determinants and evaluation: A biomechanical review and critique," *PeerJ*, vol. 7, e6364, 2019.

[4] Y. Yang and D. Ramanan, "Articulated Human Detection with Flexible Mixtures of Parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878-2890, 2013.

[5] J. C.-W. Cheung, E. W.-C. Tam, A. H.-Y. Mak, T. T.-C. Chan, W. P.-Y. Lai, and Y.-P. Zheng, "Night-time monitoring system (eNightLog) for elderly wandering behavior," *Sensors*, vol. 21, no. 3, p. 704, 2021.

[6] J. C.-W. Cheung, E. W.-C. Tam, A. H.-Y. Mak, T. T.-C. Chan, and Y.-P. Zheng, "A night-time monitoring system (eNightLog) to prevent elderly wandering in hostels: A three-month field study," *International journal of environmental research and public health*, vol. 19, no. 4, 2103, 2022.

[7] C. Picard-Deland, T. Nielsen, and M. Carr, "Dreaming of the sleep lab," *PloS one*, vol. 16, no. 10, e0257738, 2021.

[8] H. Yoon *et al.*, "Estimation of sleep posture using a patch-type accelerometer based device," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 4942-4945.

[9] M. Borazio and K. Van Laerhoven, "Combining wearable and environmental sensing into an unobtrusive tool for long-term sleep studies," in *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*, Miami, United States, 28 - 30 Jan 2012, pp. 71-80.

[10] Y. Y. Li, Y. J. Lei, L. C. L. Chen, and Y. P. Hung, "Sleep posture classification with multi-stream CNN using vertical distance map," in *2018 International Workshop on Advanced Image Technology (IWAIT)*, Chiang Mai, Thailand, 7-9 Jan 2018, pp. 1-4.

[11] J. C.-W. Cheung, B. P.-H. So, K. H. M. Ho, D. W.-C. Wong, A. H.-F. Lam, and D. S. K. Cheung, "Wrist accelerometry for monitoring dementia agitation behaviour in clinical settings: A scoping review," *Frontiers in Psychiatry*, vol. 13, p. 913213, 2022.

[12] A. Ren, B. Dong, X. Lv, T. Zhu, F. Hu, and X. Yang, "A non-contact sleep posture sensing strategy considering three dimensional human body models," in *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, Chengdu, China, 14 - 17 Oct 2016, pp. 414-417.

[13] T. Grimm, M. Martinez, A. Benz, and R. Stiefelhagen, "Sleep position classification from a depth camera using bed aligned maps," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Cancun, Mexico, 4 - 8 Dec 2016, pp. 319-324.

[14] A. Y.-C. Tam, B. P.-H. So, T. T.-C. Chan, A. K.-Y. Cheung, D. W.-C. Wong, and J. C.-W. Cheung, "A blanket accommodative sleep posture classification system using an infrared depth camera: A deep learning approach with synthetic augmentation of blanket conditions," *Sensors*, vol. 21, no. 16, 5553, 2021.

[15] A. Y.-C. Tam *et al.*, "Depth-Camera-Based Under-Blanket Sleep Posture Classification Using Anatomical Landmark-Guided Deep Learning Model," *International Journal of Environmental Research and Public Health*, vol. 19, no. 20, 13491, 2022.

[16] S. Yue, Y. Yang, H. Wang, H. Rahul, and D. Katabi, "BodyCompass: Monitoring sleep posture with wireless signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, p. 66, 2020.

[17] Y.-Y. Li, S.-J. Wang, and Y.-P. Hung, "A Vision-Based System for In-Sleep Upper-Body and Head Pose Classification," *Sensors*, vol. 22, no. 5, p. 2014, 2022.

[18] S. M. Mohammadi, S. Kouchaki, S. Sanei, D. J. Dijk, A. Hilton, and K. Wells, "Tensor Factorisation and Transfer Learning for Sleep Pose Detection," in *2019 27th European Signal Processing Conference*

(EUSIPCO), A Coruna, Spain, 2-6 Sep 2019, pp. 1-5.

[19] S. Liu, Y. Yin, and S. Ostadabbas, "In-Bed Pose Estimation: Deep Learning With Shallow Dataset," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 14, no. 7, p. 490012, 2019.

[20] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, Zurich, Switzerland, 6 - 12 Sep 2014, pp. 740-755.

[21] Z. Cao, G. Martinez, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 172-186, 2021.

[22] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation," *arXiv:2209.08575*, 2022.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, United States, 27 - 30 Jun 2016, pp. 770-778.

[24] J. Wang *et al.*, "Applications of Deep Learning Models on the Medical Images of Osteonecrosis of the Femoral Head (ONFH): A Comprehensive Review," *IEEE Access*, vol. 12, pp. 57613-57632, 2024.

[25] Z. Geng, M.-H. Guo, H. Chen, X. Li, K. Wei, and Z. Lin, "Is Attention Better Than Matrix Decomposition?," *arXiv:2109.04553*, 2021.

[26] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollar, "Designing Network Design Spaces," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, United States, 13 - 19 Jun 2020, pp. 10425-10433.

[27] S. Jin *et al.*, "Whole-Body Human Pose Estimation in the Wild," *arXiv:2007.11858*, 2007.

[28] B. Xiao, H. Wu, and Y. Wei, "Simple Baselines for Human Pose Estimation and Tracking," in *European Conference on Computer Vision*, Munich, Germany, 8-14 Sep 2018, pp. 472-487.

[29] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT Pre-Training of Image Transformers," *arXiv:2106.08254*, 2021.

[30] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," *arXiv:2105.15203*, 2021.

[31] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose++: Vision Transformer for Generic Body Pose Estimation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 46, no. 02, pp. 1212-1230, 2024.

[32] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, United States, 18 - 23 Jun 2018, pp. 7132-7141.

[33] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *European Conference on Computer Vision*, Munich, Germany, 8 - 14 Sep 2018, pp. 3-19.

[34] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks," in *38th International Conference on Machine Learning*, Online, 18 - 24 Jul 2021, pp. 11863-11874.

[35] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual Attention Network," *arXiv:2202.0974*, 2022.

[36] T. Wei, Y. Tian, Y. Wang, Y. Liang, and C. W. Chen, "Optimized separable convolution: Yet another efficient convolution operator," *AI Open*, vol. 3, pp. 162-171, 2022.

[37] L. Sifre and S. Mallat, "Rigid-Motion Scattering for Texture Classification," *arXiv.1403.1687*, 2014.

[38] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, United States, 27 - 30 Jun 2016, pp. 2818-2826.

[39] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," *arXiv:1706.05587*, 2017.

[40] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," *arXiv:1511.07122*, pp. 240-245, 04/30 2016.

[41] P. Wang *et al.*, "Understanding Convolution for Semantic Segmentation," *arXiv:1702.08502* 2017.

[42] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded Pyramid Network for Multi-Person Pose Estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018 2018, pp. 7103-7112.

[43] D. K.-H. Lai *et al.*, "Dual ultra-wideband (UWB) radar-based sleep posture recognition system: Towards ubiquitous sleep monitoring," *Engineered Regeneration*, vol. 4, no. 1, pp. 36-43, 2023.

[44] D. K.-H. Lai *et al.*, "Vision Transformers (ViT) for Blanket-Penetrating Sleep Posture Recognition Using a Triple Ultra-Wideband (UWB) Radar System," *Sensors*, vol. 23, p. 2475, 2023.

[45] K. Hou, S. Xia, and X. Jiang, "BuMA: Non-Intrusive Breathing Detection using Microphone Array," in *MobiSys '22: The 20th Annual International Conference on Mobile Systems, Applications and Services*, Portland, United States, 27 Jun 2022, pp. 1-6.