Smart Multimodal In-Bed Pose Estimation Framework Incorporating Generative Adversarial Neural Network

Sumit Singh, Mohammad Hossein Anisi, *Senior Member, IEEE,*, Anish Jindal, *Member, IEEE* and Delaram Jarchi, *Senior Member, IEEE*

***Abstract*—Monitoring in-bed pose estimation based on the Internet of Medical Things (IoMT) and ambient technology has a significant impact on many applications such as sleep-related disorders including obstructive sleep apnea syndrome, assessment of sleep quality, and health risk of pressure ulcers. In this research, a new multimodal in-bed pose estimation has been proposed using a deep learning framework. The Simultaneously-collected multimodal Lying Pose (SLP) dataset has been used for performance evaluation of the proposed framework where two modalities including long wave infrared (LWIR) and depth images are used to train the proposed model. The main contribution of this research is the feature fusion network and the use of a generative model to generate RGB images having similar poses to other modalities (LWIR/depth). The inclusion of a generative model helps to improve the overall accuracy of the pose estimation algorithm. Moreover, the method can be generalized for situations to recover human pose both in home and hospital settings under various cover thickness levels. The proposed model is compared with other fusion-based models and shows an improved performance of 97.8% at PCKh@0.5. In addition, performance has been evaluated for different cover conditions, and under home and hospital environments which present improvements using our proposed model.**

***Index Terms*—Internet of Medical Things, AI, SLP, Generative adversarial neural network, LWIR, depth.**

I. Introduction

In-bed pose estimation can monitor the quality of sleep which is vital in several healthcare prognosis, diagnosis, and therapy practices. According to research by [1], a poor in-bed sleeping posture increases the chance of several medical conditions like carpal tunnel syndrome, sleep apnea, and pressure ulceration. Caregivers are extensively employed for the estimation of in-bed postures based on visual inspection; however, this process is tiresome, and the readings and evaluation techniques are subjective [2]. Wearable gadgets are employed for tracking the quality and posture during sleep, but these devices are intrusive in nature [3]. The advancements in computer vision have empowered camera-based contactless in-bed pose estimation techniques [4]. These methods require less maintenance, are less expensive, and are comfortable for patients. Convolutional pose machines [5] have led the foundation for several 2D and 3D human pose estimation methods [6] [7]. However, the accuracy of in-bed pose estimation by the state-of-the-art (SOTA) method is hindered due to the presence of heavy occlusion and extreme illumination [8].

Recent literature on pose estimation emphasises on skeleton-based top-down methods because of their flexible and easy representation [9]. Performance of SOTA methods [10] [11] [12] for pose estimation is enhanced by adding a multi-scale feature processing technique, where the features at various scales are used in parallel. Over the last few years, the high-resolution net (HRNet) [11] has gained substantial popularity as not only it attains efficient training and inference speed but also preserves high-quality features across the network. Therefore, it has shown an enhanced localised accuracy for pose estimation. Fine-tuning methods for in-bed pose estimation have found several healthcare applications such as fall risk assessment [13] and seizure disorder classification [14]. These methods have showcased the potential impact and versatility of in-bed human pose estimation in addressing crucial healthcare challenges.

However, several challenges for developing in-bed human pose monitoring algorithms still exist such as the selection of the best data recording modality/modalities and a reliable benchmark for validation purposes considering real-world environments. The Simultaneously-Collected Multimodal Lying Pose Dataset (SLP) dataset [15] has been recently released publicly where a large-scale benchmark was created including various imaging modalities. The SLP dataset [15] consists of simultaneous imagery from four modalities: Visual image (RGB), Long wavelength Infrared (LWIR), depth and pressure maps (PM). Moreover, in [15] a number of SOTA 2D human pose estimation algorithms are evaluated for quantifying the performance of in-bed human pose monitoring using the SLP dataset. These algorithms are evaluated using a single modality. Selection of the most favourable modality for pose estimation might be difficult as RGB images become inefficient under occluded and illuminated conditions. Using LWIR, it is possible to see under the cover but residual heat from the body affects the thermal imagery. PMs are not affected by residual heat, but it is difficult to recognise arms and heel joints due to the lack of pressure concentration regions. Depth might reconstruct an object's geometry in illuminated and occluded conditions, but they are affected by different elevated areas due to specific in-bed poses. Moreover, LWIR

S. Singh, M. H. Anisi, and D. Jarchi are with the School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, United Kingdom (e-mails: ss20727@essex.ac.uk; m.anisi@essex.ac.uk; delaram.jarchi@essex.ac.uk)

A. Jindal is with the Department of Computer Science, Durham University, Durham DH1 3LE, United Kingdom (e-mail: anish.jindal@durham.ac.uk).

and depth modalities are not as expensive as the PMs modality. In-bed human pose estimation is challenging due to (i) the presence of heavy occlusion (thin and thick cover), (ii) lack of luminated environment, (iii) privacy concerns of the patients, and (iv) expensive imaging tools and sensors. In recent studies, contactless imaging such as using depth [16] and LWIR [17] modalities have been used to infer contact pressure [17], [16] as PMs are expensive and these studies found them not suitable for continuous monitoring and everyday practice [16], [17].

The realistic way to overcome the modality selection issue is by developing an effective pose estimation technique that employs a feature fusion from privacy-protecting and inexpensive modalities such as LWIR and depth. These modalities do not capture distinctive or identifying characteristics related to face, clothing, or surrounding environment [18]. In such cases, additional privacy-preservation measures are not required contrary to the use of RGB which contains privacy-related data and therefore, additional methods such as visual privacy preservation techniques are required [19]. However, more privacy can be preserved during data collection where an inference model can be conducted on-the-chip and there is no need to save or transfer any image to third-party sources [18]. Therefore, LWIR and depth modalities are selected in our research to train our proposed method for improving the accuracy of pose estimation. Moreover, we use a generative model to synthetically generate uncovered RGB images as they are ideal for pose estimation.

We propose a deep learning-based IoMT network architecture for robust and accurate in-bed human pose estimation. The proposed architecture is aimed at providing IoMT solutions such as for remote in-bed pose monitoring for patients using contactless technology. These include monitoring patients with sleep apnea and pressure ulcers which can be extended for continuous monitoring of infants' pose in the neonatal intensive care unit for timely and effective treatments [20], [15]. Our network consists of four phases: 1. **Pre-processing**, where each modality is aligned, the region of interest (ROI) is selected and the image is enhanced by the histogram equalisation method. 2. **Feature fusion** of the pre-processed modalities (LWIR and depth) by an autoencoder and fusion module such that the modalities are fused without any variational patterns (such as any different elevated regions by depth modality and/or different heat patterns over blanket for LWIR). 3. **Synthetic image generation** of uncovered RGB image from fused modality by proposing and employing a generative adversarial network (GAN)-based model. The generator is based on the U-Net structure, and the initial image is down-sampled for a residual block whose output is fused with the penultimate transpose convolutional layer. We have generated uncovered RGB image from fused modality because pose estimation algorithms seem to work best with uncovered RGB images. 4. **Pose estimation** by modified HRNet which can preserve global spatial information and draw the pose skeleton with much higher accuracy. The proposed method is rigorously tested over SLP dataset against the SOTA methods, where it outperforms on all considered evaluation parameters.

Our contribution:

1) We propose a feature fusion technique based on autoencoder architecture with dense convolutional blocks and a proposed fusion module. It concatenates the deep features from two modalities (LWIR and depth) and produces a fused feature image.
2) We introduce a novel architecture for the generation of synthetic image from one domain to another. Our generator model includes a two-stage image generation module which preserves the global features and increases the quality of the synthetic image.
3) We modify the architecture of HRNet to achieve higher accuracy for pose estimation. A spatial pyramid pooling layer along with an attention mechanism is added to preserve spatial information and promote the reuse of features.
4) We demonstrate the effectiveness of our proposed method and how it increases the performance of in-bed human pose estimation on SLP dataset with 109 adult subjects in different environments (home and hospital) covered under the different thicknesses of the sheet.

The remainder of the paper is structured as follows. In Section II the methods including network architecture are described. This section provides details on pre-processing algorithms, feature fusion techniques, and synthetic image generation. In Section III, experimental results are explained including implementation details, SLP dataset, and evaluation metrics. Finally, the paper is concluded in Section IV with a discussion on further work.

II. Methodology

In this section, we describe pre-processing techniques applied to different modalities, then, feature fusion, generation of syntactic images, and human pose estimation method are explained. The complete architecture of our proposed method is represented in Fig 1. Each component of the proposed method is discussed in detail in this section. More specifically, Section II.A outlines the elementary architecture of the proposed technique. In Section II.B, the pre-processing approach is described. Section II.C introduces the feature fusion technique whereas the method for generation of syntactic RGB images is described in Section II.D. Finally, the modified pose estimation algorithm is presented in Section II.E.

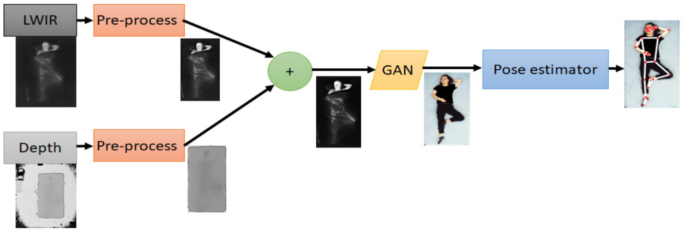

Fig. 1. Proposed architecture for in-bed human pose estimation.

A. Network Architecture

In this research, a multi-staged human pose estimation architecture is designed. We have selected two modalities (i.e., LWIR and depth) from the SLP dataset for in-bed pose estimation because these modalities are not only privacy protecting and inexpensive but also it has been shown that they perform significantly well under non-illuminated and occulated environments. The images from both of these modalities are pre-processed before fussing the features using a proposed encoder-decoder network. Thereafter, the fused modality is processed by the GAN model which has been customised to generate RGB images of patients without the cover in a similar pose as in the LWIR/depth modality. The generation of uncovered synthetic images increases the accuracy of pose estimation as uncovered and illuminated RGB images can be trained well with pose estimation algorithms.

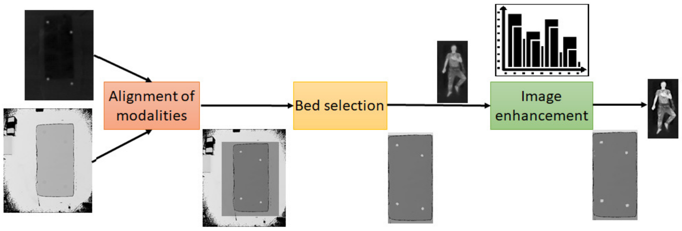

Fig. 2. The overview of pre-processing of modalities which includes three stages: Alignment of modalities, Bed selection, and Image enhancement.

B. Pre-processing

In this section, pre-processing stages are outlined where we have employed three distinct methods: 1. Modality alignment, 2. Bed selection, and 3. Image enhancement. Fig. 2 represents a visual demonstration of the pre-processing steps. The LWIR and depth modalities possess dissimilar resolutions, as 120×160 and 424×512, respectively. Due to this dissimilarity, fusing features from these modalities will be challenging. Hence, the initial step involves aligning the images based on a marker present in the SLP dataset. The SLP dataset consists of four markers situated on the bed, each with a fixed height of 10cm. These markers possess heat-emitting capabilities and are characterized by a fixed weight and a distinct red coloration on the top surface. These features help in the calibration of the markers using all four modalities to obtain: RGB, LWIR, PM, and depth. Subsequently, the images are calibrated by referencing the markers on each modality. Following this calibration process, a region of interest (I_{ROI}) is calculated based on the markers position. This selection not only improves computational efficiency but also directs the subsequent pose estimation methods to concentrate exclusively on the I_{ROI}. Henceforth the modalities are resized to 128×256 pixels such that effective training of deep learning can be achieved without loss of any key features. Finally, the histogram-based image enhancement method is employed to rectify the contrast and brightness of the image to ensure that each modality can be enhanced and luminated. The pre-processing phase seems to be enhancing the results of feature fusion and thereafter in-bed pose estimation.

C. Feature fusion

This section provides a comprehensive introduction to the proposed fusion method based on deep learning. The input LWIR and depth modalities are represented as I_{LWIR} and I_{Depth}. The proposed architecture comprises of three key components: 1. Encoder, 2. Fusion module and 3. Decoder. Fig. 3 illustrates the architecture of the proposed network.

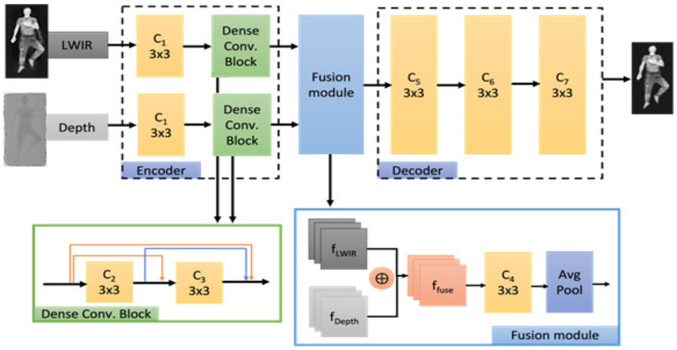

Fig. 3. A detailed framework of proposed deep learning-based fusion method.

The encoder block has two modality specific feature extraction (MSFE) modules due to the variability of components in each modality. MSFE allows the modalities to process independently, thereby allowing the network to capture modality-specific patterns and information more efficiently. It also allows the network to focus on specific characteristics and exploit their complementary information. Each MSFE module consists of a convolutional layer and a dense convolutional block for extraction of deep features of modalities. Generalised features are extracted by the first convolutional layer ($C1$) with a filter size of 3×3, thereafter a dense convolutional block (where the input of each layer is the cascaded input and output of the previous layer) with two convolutional layers ($C2$ and $C3$) of 3×3 filter size is employed to capture deep features. The dense connections in the block facilitate the flow of information within the modality-specific paths, allowing for iterative and enriched feature learning. This iterative refinement helps to capture finer details and improve the discriminative power of the features. Each convolutional layer in the encoder has a filter size of 3×3 and a stride of 1. This approach ensures the use of salient features in the fusion module, and it also helps the deep features to be preserved in the encoder.

The two feature maps received by the encoder are concatenated in the fusion module. Concatenation is a widely used technique for feature fusion as it preserves individual modality-specific information while enabling the network to leverage the combined knowledge from both modalities. There is n number of feature maps for each modality represented as, $n \in 1, 2, \ldots, N$, $N = 48$. I_{LWIR} and I_{Depth} are the two-input modality for the encoder that are pre-processed and resized. Next, the encoder produces feature maps of ϕ^n_{LWIR} and ϕ^n_{Depth}, from pre-processed images of both modalities which then pass through three convolutional layers of $C1$, $C2$ and $C3$. f^{fused} denotes the fused feature maps. The corresponding point in the input and fused feature maps is (i, j). The concatenation strategy is formulated by the following equation:

$$f^{fused}(i, j) = \phi^n_{LWIR} \oplus \phi^n_{Depth} \tag{1}$$

By concatenating the feature maps, the fusion module enables the network to access information from both modalities simultaneously. The concatenated feature maps contain a joint representation that incorporates information from both modalities (LWIR and depth), allowing for the integration of the complementary features and facilitating cross-modal interactions. Thereafter, the fused feature maps are further processed through convolutional layer (C_4) and pooling operation to capture higher-level representations and refine the fused features. The output of the pooling layer serves as the input to the decoder, which is responsible for the reconstruction of fused images. The decoder block consists of three convolutional layers of filter size 3×3 and stride 1. A simple yet efficient model is used for the reconstruction of fused image.

A detailed model summary of the proposed architecture is outlined in Table I. The encoder block has one convolutional layer (C_1) with a filter size of 3×3, whereas the dense convolutional block has two convolutional layers, C_2 and C_3, which are connected by cascading operation to catch dense features. Rectified Linear unit (ReLu) is used as the activation function because of its computational efficiency. The fusion block has a convolutional layer followed by an average pooling layer of filter size of 3×3. The decoder block has three convolutional layers: C_5, C_6, and C_7 which are used to reconstruct the fused feature image. We have employed the Programmatic Rectified Linear unit (PReLu) in the decoder block as it has the functionality to learn parameters to control the slope of the negative part of the activation function. By assigning negative values to create different slopes, PReLU can adaptively capture more complex and diverse representations. The proposed model is trained using the SLP dataset with LWIR and depth modalities. The model is trained for 300 epochs with a batch size of 16 and a learning rate of 1×10^{-5}. The encoder-decoder network is trained to yield enhanced results for feature extraction and regeneration of fused image.

TABLE I
FEATURE FUSION SPECIFICATIONS FOR PROPOSED ARCHITECTURE

Block	Layers	Input	Output	Activation	Filter	Stride
Encoder	C1 Dense Conv.	1	16	ReLu	3 × 3	1 1
Dense	C2	16	16	Leaky ReLu	3 × 3	1
conv	C3	32	16	Leaky ReLu	3 × 3	1
Fusion	C4 Avg. pool	48 48	48 32	ReLu ReLu		1
Decoder	C5 C6 C7	32 32 16	32 16 1	PReLu PReLu	3 × 3 3 × 3 3 × 3	1 1 1

D. Synthetic image generation

In this section, a privacy-preserving method of synthetic image generation from the fused modality (LWIR + Depth) is proposed. Our technique can generate RGB image of uncovered human-on-bed from fused modalities. The synthetically generated visible images do increase the accuracy of in-bed human pose estimation as uncovered RGB images do provide higher scores for pose estimation when compared with other covered images from various modalities. Fig. 4 shows the pictorial representation of the proposed architecture. We have explored the architecture of GAN proposed by [21] and adopted by many researchers across the globe for many different applications. Moreover, GAN has been used widely for the generation of high-resolution images such as RGB.

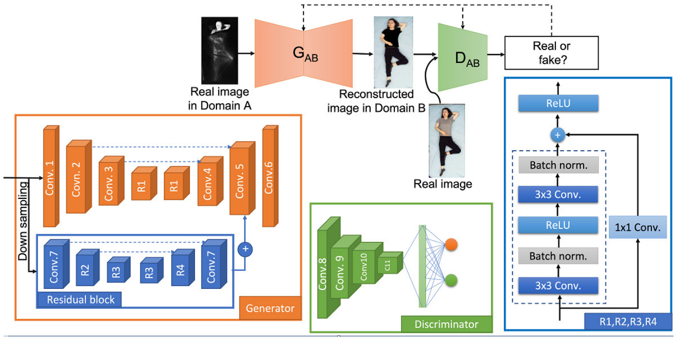

Fig. 4. Network architecture of proposed GAN model.

Random noise vector z is used for the generation of meaningful output image y, $G : z \rightarrow y$. The output generated by the generator (G) tries to be indistinguishable from the original image, whereas a discriminator (D) is trained to identify generated images and penalise the generator block such that G can generate more realistic images. The task of D is to effectively differentiate real or fake images. The fused image is taken as input by the generator to reconstruct the uncovered RGB image and the discriminator block is calculated as the difference between the corresponding real image and the reconstructed image.

The generator block is inspired by the architecture of UNet [22] which included encoder and decoder blocks with skip connections being established between corresponding layers. These connections help to preserve and transfer low-level features from the encoder to the decoder, facilitating the generation of fine details in the output image. It also has the residual layer (R1) which captures more complex and abstract features thereby leading to improved performance for image translation from one domain to another. In this type of architecture, the input undergoes a sequence of down-sampling layers until reaching a bottleneck layer, after which the process is reversed. This design ensures that all information must pass through every layer, including the bottleneck. However, the image translation tasks have common low-level information between input and output which would be advantageous to directly transfer this shared information across the network.

Thus, adding a skip connection in an encoder-decoder architecture is important. A residual block is added to the generator architecture to capture the global features of the fused modality. The input image is down-sampled by 2× such that the global features like edge and texture of the in-bed human are captured. This block contains (i) a convolutional layer for extracting initial key features and encoding spatial information, (ii) a series of residual layers for preserving low-level details and learning residual information, and (iii) a transpose convolutional layer for up-sampling and adjusting channel dimension. These residual blocks help in overcoming the vanishing gradient problem and capture the deep features. Thereafter, the result is concatenated with Conv.5 of the main architecture such that the local and global features can be fused. The generator architecture shows an efficient flow of information with skin connections for feature retention and parameter sharing which enables faster inference. The feed-forward structure allows for fast processing of input data, thus enabling faster pose estimation. Each residual layer consists of a series of convolutional, batch normalisation, and ReLU layers along with a 1 × 1 convolution layer used for the transformation of feature maps. This architecture helps in feature extraction and translates the preserved features to the next layers.

The discriminator architecture is a sequence of convolutional neural network with a fully connected layer at the end. Such architecture provided spatial awareness and helped in localized discrimination. It also reduces the computational complexities and increases the robustness to changes in global features thereby improving the training stability of the discriminator block. These characteristics make it well-suited for image translation tasks, where preserving local details and maintaining realistic and coherent patches are crucial for generating high-quality outputs. The model summary for both the generator and discriminator block is tabulated in Table II, where the number of the input channel (I/p), filter size, stride, and activation are mentioned for each layer.

TABLE II
PROPOSED ARCHITECTURE SPECIFICATIONS FOR GAN MODEL

Layers	I/p	Filter	Stride	Activation
Generator				
Conv.1	32	5 × 5	1	ReLU
Conv.2	64	3 × 3	2	ReLU
Conv.3	128	3 × 3	2	ReLU
R1	128	3 × 3	1	LeakyReLU
Conv.4	128	3 × 3	2	LeakyReLU
Conv.5	64	3 × 3	2	ReLU
Conv.6	3	5 × 5	1	tanh
Residual Block				
Conv.7	64	3 × 3	2	ReLU
R2	128	3 × 3	2	ReLU
R3	256	3 × 3	2	ReLU
R4	128	3 × 3	2	ReLU
Discriminator				
Conv.8	32	3 × 3	2	LeakyReLU
Conv.9	64	3 × 3	2	LeakyReLU
Conv.10	128	3 × 3	2	LeakyReLU
Conv.11	256	3 × 3	2	LeakyReLU

For conditional GAN, the reconstructed synthetic image (I_{gen}) by the generator $G(I_{fused}, z)$ is conditioned on input I_{fused} and is compared against the corresponding conditioned real image (I_{RGB}) in the discriminator (D). The objective function is for conditional GAN is shown in Equation (2):

$$\mathcal{L}_{cGAN}(G, D) = \Xi_{I_{fused}, I_{RGB}}[logD(I_{fused}, I_{RGB})] + \Xi_{I_{fused}, z}[log(1 - D(I_{fused}, G(I_{RGB}, z)))] \quad (2)$$

The fundamental objective of the generator is to maximise the discriminator loss and minimise the generator loss by fooling the discriminator by generating high-resolution RGB images. We have used the $L2$ loss function as it penalizes errors more heavily due to squaring operations, thereby making it more sensitive to outliers and minimising the loss between I_{gen} and I_{RGB}. It tends to construct smoother results as it penalizes large errors more and encourages predictions that are closer to the mean. The objective function of the method and the loss are represented in Equations (3) and (4) respectively:

$$G = argmin_G max_D \mathcal{L}_{GAN}(G, D) + \lambda \mathcal{L}_{L_2}(G) \quad (3)$$

$$\mathcal{L}_{L_2}(G) = \Xi_{I_{fused}, I_{RGB}}[(I_{RGB} - G(I_{fused}, z))^2] \quad (4)$$

GANs have the ability to translate images from one domain to another by mapping input to output, but we have added gaussian noise (z) along with I_{fused} in G, to avoid deterministic outputs. [23] have acknowledged that adding noise along with input to the generator provides better results. The standard approach of training a GAN module is employed for our model, which was proposed by [21] where one-step gradient descent (SGD) on D and G are trained alternatively to maximize log D $(I_{RGB}, G$ $(I_{fused}, z))$. Additionally, we have also slowed the learning rate of D such that G can be trained without vanishing its gradients. A learning rate of 0.0001 is used along with SGD optimiser to train our model.

E. Pose estimation

We have adopted the architecture of HRNet [11] and have modified it by adding a convolutional layer and a Spatial Pyramidal Pooling Block (SPPB) prior to the original structure of HRNet. Additionally, we have also added a dense block as an attention mechanism between stage 2 and stage 3 of HRNet. The modified architecture increases the accuracy of in-bed human pose estimation from synthetically generated image. The flowchart of the modified pose estimation network is pictorially represented in Fig. 5.

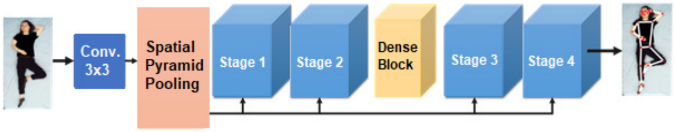

Fig. 5. Network architecture of pose estimation module.

A convolutional layer of filter size 3 × 3 is added before the SPPB such that it can further enhance the representation of features before performing multi-scale pooling. The network can learn more discriminative and abstract features, which can then be pooled and processed at multiple scales by the SPPB. This combination enhances the model's ability to capture both local and global contextual information, leading to improved performance in pose estimation tasks. The Feature Pyramid Networks (FPN) [24] inspires the use of the feature pyramid, which is extensively employed by the architecture of object detection modules. It leverages the intrinsic multi-level features to provide rich semantic knowledge at all levels. The pyramid structure follows a top-down architecture with four levels, where convolution stride is employed for sampling down the features. Each down-sampling step reduces the spatial resolutions (height and width) of the feature maps by half while doubling the feature dimensions (number of channels). Within this hierarchical structure, the primary high-resolution branch of the network is directly fed with the top-level feature map with the highest resolution. Meanwhile, the subsequent levels (2^{nd}, 3^{rd}, and 4^{th}) of feature maps are merged with their respective peer by the process of element-wise addition from the multi-resolution branches. This fusion process combines the finer details from the higher-resolution maps with the broader context captured by the lower-resolution maps, resulting in a comprehensive representation at each level of the feature pyramid. Across various semantic levels within our network architecture, we have merged the feature maps with corresponding peers by incorporating lateral connection at the beginning of each parallel branch such that we can take advantage of the feature pyramid's output at each level. The mathematical representation of the fusion of features at multi-level is expressed by the following equation:

$$F_n = I_n + H_{n-1} \tag{5}$$

The symbol F_n represents the output of the fusion operation, where the feature map from the k^{th} level of the feature pyramid, denoted as I_n, is element-wise added to its corresponding counterpart H_{n-1} from the $(n-1)^{th}$ parallel branch in the network. The parameter n takes on values 2, 3, and 4. Notably, the 1st branch corresponds to the primary branch in the architecture, and the top-level feature map of the pyramid directly feeds into this branch.

After the multi-level feature fusion, F_n contains enriched semantic information derived from the feature pyramid. This fused feature map becomes the new input for the n^{th} parallel branch, starting the subsequent processing at that level. By incorporating SPPB in front of the HRNet architecture, we have harnessed its benefits to improve scale invariance, capture spatial context, achieve computational efficiency, and preserve spatial information. These advantages contribute to more robust and accurate pose estimation, particularly in challenging scenarios. It helps the HRNet architecture gain a more comprehensive understanding of the spatial context surrounding each body joint. This enhanced spatial context can aid in accurate joint localization and pose estimation, especially in situations where contextual information is critical, such as complex in-bed human poses.

A dense block is added as an attention mechanism between stage 2 and stage 3 of the HRNet architecture, it potentially increases the accuracy of in-bed human pose estimation. The Dense Block (DB), originally introduced in the Dense Net architecture [25], incorporates dense connections between layers, allowing each layer to receive feature maps from all preceding layers. This design promotes feature reuse and facilitates information flow throughout the network. DB consists of three convolutional layers (of filter size 3×3), each followed by a concatenation operation to combine the feature maps from all preceding layers. Batch normalization and ReLU as an activation function are applied to each layer. This promotes the extraction of meaningful features and helps in attending to relevant information for pose estimation.

III. Results and Discussions

In this section, we will discuss the implementation details, the dataset used for training and testing our model, and the performance metrics to evaluate the performance of our method. Moreover, the proposed pipeline is compared with SOTA to evaluate the effectiveness of the model.

A. Implementation Details

Our network architecture is trained on NVIDIA V100 GPU and TensorFlow along with Python is used to design the network. The feature fusion network is trained for 100 epochs while the GAN and pose estimation network is trained for 200 epochs. The convolutional weights are initialized to a normal distribution with a standard deviation of 0.05 and a mean of 0. Adam optimization algorithm with β_1= 0.5 and β_1= 0.1 is used with the constant learning rate as 1×10^{-5} by the feature fusion model. Additionally, for the GAN model, we have also slowed the learning rate of D such that G can be trained without vanishing its gradients. A learning rate of 0.0001 is used along with SGD optimiser to train our model. The pose estimation architecture is then trained by the Adam optimizer with an initial learning rate of 0.0002 and a decay rate of 0.01 for the next 100 epochs.

B. SLP Dataset

The proposed pipeline is trained and tested on publicly available SLP dataset [15]. The dataset has in-bed human pose images taken from four different modalities (RGB, LWIR, Depth, PM) under three cover conditions (no cover, ∼1mm thin cover, ∼3mm thick cover) for 109 participants at home and hospital environmental conditions (Dana Lab and Sim Lab respectively). Additionally, the dataset records 45 different natural sleeping poses for all 109 participants under 3 main sleeping posture categories: right side, supine, and left side. The images are captured in difficult luminance and occultation conditions by all the modalities; thus, it helps the deep learning model train under natural sleeping conditions (occulted by the cover and under low luminance). Moreover, to increase the variability of the dataset, different colour covers are used for home and hospital settings. Samples of dataset under different cover conditions by different modalities are shown in Fig. 6. For visual representation the image for PM in Figure Fig. 6 is changed to a blue channel such that the image is visible. The image resolution for each modality is different, RGB image is of 576×1024, LWIR is 120×160, Depth is 424×512 and the PM is 84×192. This variance in image resolution also affects the training process so we have pre-processed the modalities before using it in our deep network architecture. The dataset also contains the annotated ground truth (x, y) coordinates for fourteen body joints. Table III tabulates the train-test split of data under different environment settings. As shown in the Table, for the home environment, there are 102 subjects considering 45 frames and 3 different cover conditions, thereby, a total of $102 \times 45 \times 3 = 13770$ samples. Out of these samples, 12150 samples have been used for training and 1620 samples have been used for the

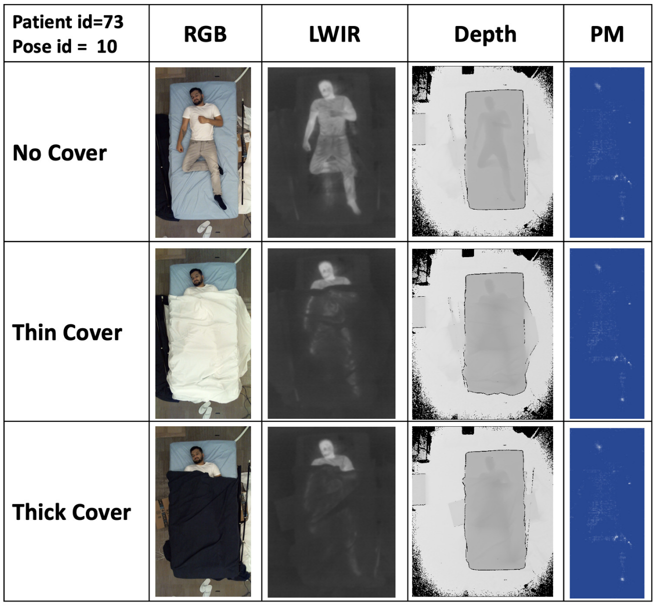

Fig. 6. Qualitative review of SLP dataset for patient id: 73 and pose id: 10.

testing. The total number of testing samples considering the hospital environment is 2565 samples. It is worth noting that these samples correspond to the number of samples for each modality, i.e., one modality.

TABLE III
OVERVIEW OF SLP DATASET

Environment	Subject(train+test)	Train	Test
Home	102(90+12)	12150	1620
Hospital	7(0+7)	0	945
total	109	12150	2565

We have used LWIR and depth modalities for training and testing our proposed pipeline as they are inexpensive unlike the PM modality and protect the privacy of the sleeping patients, unlike RGB modality. The pose estimation algorithm on the selected modalities performs better than the other two as it is not affected by occultation or luminance as the other two modalities. Moreover, it is also noticed that PM modality results in a ghost effect and fails to locate heels and arms thereby making it difficult for pose estimation under certain sleeping postures.

C. Evaluation Metrics

We evaluated the performance of synthetic image generation by the GAN network on three evaluation metrics: Frechet Inception Distance (FID) [26], Peak Signal-to-Noise Ratio (PSNR) [27] and Structural similarity Metric (SSIM) [28].

FID measures the distance between generated and real image at the feature level, it is mathematically represented as:

$$FID = \|M_{real} - M_{generated}\|^2 + \\ Tr(C_{real} + C_{generated} - 2\sqrt{(C_{real} \times C_{generated})}) \qquad (6)$$

Where M_{real}, $M_{generated}$ are the mean of feature vectors of real and generated image respectively. The covariance of the real and generated image are represented as C_{real}, $C_{generated}$, and T_r is the trace of the matrix.

PSNR measures the peak error between real and generated image. It calculates the difference between images based on pixel values. A higher value of PSNR shows a low perpetual difference between images. The mathematical representation of PSNR is:

$$PSNR = 20 \times log_{10} I_{max} - 10 \times log_{10} MSE \qquad (7)$$

Where, I_{max} is the maximum possible pixel value and MSE is the mean square error between the real and generated image. The similarity between the real and generated image in terms of brightness, contrast, and structure is measured by SSIM. It is designed to mimic human perception on how we visualise images rather than pixel-wise comparison. It is represented as:

$$SSIM(i,j) = L(i,j) \times C(i,j) \times s(i,j) \qquad (8)$$

where $L(i,j)$, $C(i,j)$, $s(i,j)$ are the similarity components for Luminance, Contrast, and Structure, respectively. The value of SSIM lies between 0 to 1 where 1 is associated with the generated image that exactly looks like the corresponding real image. We have used a percentage of correct keypoint at normalised distance of 0.5 (PCKh@0.5 metric) to assess the performance of pose estimation framework. It is the measure of distance between ground truth and predicted joint at less than 50% of head bone length. The distance between the points representing thorax and head is used to calculate head bone length. The effectiveness of the proposed pipeline for pose estimation is validated on both the environment (home and hospital) using these evaluation metrics.

D. Experimental Results

The performance of in-bed pose estimation for each modality under the home setting is reported in Fig. 7. It shows that the fused modality, i.e., LWIR+depth performs better than single modalities for pose estimation for PCKh metric at 0.5. It is clear that the depth modality has a more stable performance when compared with LWIR for all three cover conditions (having PCKh@0.5 $>= 95\%$). The PM modality shows the most stable performance for all three cover conditions although its PCKh@0.5 lies the lowest among all three modalities as PM is unable to draw joints due to missing pressure points

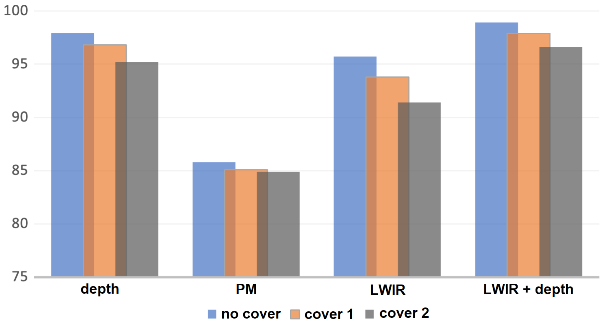

Fig. 7. PCKh performance of pose estimation of proposed architecture on depth, PM, LWIR, and fused (LWIR+depth) modalities under home setting.

at arms and heels. The results show that the performance of pose estimation is dropped under thick cover (cover 2) by LWIR modality while the PM modality is not affected much by the type of cover condition and the depth modality shows a minor difference in performance. Fusion of LWIR and depth modalities shows enhanced results for in-bed human pose estimation.

A comparison of the performance of in-bed pose estimation between the proposed method and SOTA under different environments (home or hospital setting) and different cover conditions is tabulated in Table IV. The experimental results show that the proposed method outperforms the SOTA algorithms for pose estimation under all the different cover conditions and for both the environmental conditions (home and hospital). The feature maps are merged from different scales at corresponding levels in [11], which makes it more effective for visually difficult poses in the RGB domain. By combining the high-scale feature maps, the accuracy for localization of joints is decreased but does not predict false positives. [10] uses a CNN-based self-attention network for 2D and 3D estimation pose which makes it effective for the detection of pose from RGB images.

TABLE IV
PERFORMANCE EVALUATION OF PROPOSED METHOD AND SOTA (PERCENTAGE AT PCKH@0.5) FOR DIFFERENT ENVIRONMENTAL AND COVER CONDITIONS

Dataset	Condition	Proposed method	[11]	[10]
DANA LABS (Home Setting)	No cover	98.9	98.1	97.9
	cover1	97.9	97.0	95.2
	cover2	96.6	95.9	92.5
	average	**97.8**	97.0	95.2
SIM LABS (Hospital Setting)	No cover	98.9	98.6	98.4
	cover1	97.7	97.1	97.6
	cover2	88.4	85.6	86.9
	average	**95.0**	93.8	94.3

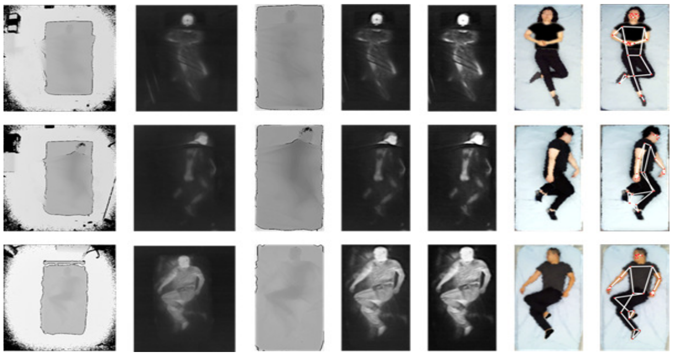

Fig. 8. Qualitative analysis of the proposed pipeline. (a) Input depth modality (b) Input LWIR modality (c) Pre-processed depth image (d) Pre-processed LWIR image (e) Feature fusion of both the modalities (f) Generation of uncovered RGB image (g) Pose estimation on synthetically generated image.

Table V shows the quantitative study of different methods in the proposed pipeline for in-bed human pose estimation using the two-privacy protecting and inexpensive modalities (LWIR and depth). We have also illustrated how the performance of [11] increases when our pipeline is employed along with the SOTA method for all the cover conditions (c0=no cover, c1=thin cover, c2=thick cover). The results show the effectiveness of the proposed architecture in improving the PCKh score @ 0.5. Table V clearly shows how the accuracy of pose estimation increases for all the cover conditions with the implementation of each stage by both methods (Proposed method and [11]). Initially, the pose estimation (PE) technique is employed on LWIR modality by both methods, and thereafter pre-processing (PP) on modality is applied, which increases the performance to a certain extent. The performance of pose estimation is spiked when the feature fusion (FF) technique is employed to the pre-processed LWIR and depth modalities. Henceforth the generation of uncovered in-bed human image by GAN further increases the performance. The performance of c2 is increased significantly as the GAN module helps the covered image of fused modality to generate a corresponding uncovered image thereby increasing the pose estimation. The pose estimation score of proposed networks for all the stages is better than the SOTA method as adding SPPB along with dense block helps the HRNet learn more spatial features thereby increasing its pose estimation capacity.

TABLE V
PERFORMANCE EVALUATION OF DIFFERENT STAGES IN PROPOSED METHOD AND SOTA (PERCENTAGE AT PCKH@0.5) UNDER DIFFERENT COVER CONDITIONS

Method	Proposed method				[11]			
Stages	c0	c1	c2	avg	c0	c1	c2	avg
PE (on LWIR)	95.4	93.3	91.8	93.5	95.2	92.6	90.9	92.9
PE+PP (on LWIR)	96.1	93.8	92.3	94.1	95.7	93.5	91.9	93.7
PE+PP+FF	97.8	96.7	95.5	96.7	97.1	95.4	94.9	95.8
PE+PP+FF+GAN	**98.9**	**97.9**	**96.6**	**97.8**	98.1	97.0	95.9	97.0

We have also evaluated the performance of the proposed GAN module for the generation of synthetic image from fused modality. The GAN model is evaluated on three parameters: FID, PSNR, and SSIM. Where the smaller value of FID is better, and the higher value of PSNR and SSIM are better. Detailed comparison of the proposed GAN model with SOTA for generation of uncovered RGB images of in-bed human is tabulated in Table VI. It shows that our model achieves better results on most parameters. Our model can not only extract the deep features from the image but also retain semantic information. [29] explored conditional adversarial architecture for image translation from one domain and other. The learning not only involves mapping of features from the corresponding input and output images but also employing the loss function to refine and train this mapping. [30] employed conditional GANs for generating high-resolution image from semantic label maps. Synthetic images are generated by U-Net architecture by [22].

TABLE VI
PERFORMANCE EVALUATION OF PROPOSED GAN MODEL WITH SOTA

Model	FID	PSNR	SSIM
[29]	231.27	8.01	0.90
[30]	217.79	**9.93**	0.91
[22]	246.14	7.34	0.86
Proposed method	**206.51**	8.68	**0.94**

We have compared the proposed pose estimation architecture with SOTA models such as [31] [11] [10] [32] [33] [34].

TABLE VII
EVALUATION OF PCKH@0.5 FOR POSE ESTIMATION UNDER DIFFERENT COVER CONDITIONS WITH DIFFERENT PRIVACY-PROTECTING MODALITIES.

Method	LWIR+PM				PM+depth				depth+LWRI			
	c0	c1	c2	avg	c0	c1	c2	avg	c0	c1	c2	avg
proposed method	95.7	93.8	91.4	93.6	91.3	89.1	83.2	87.9	**98.9**	**97.9**	**96.6**	**97.8**
[31]	95.8	93.5	92.6	93.9	90.1	90.2	90.2	90.2	97.6	96.1	95.8	96.5
[11]	95.2	93.1	91.9	93.4	84.4	84.3	84.2	84.3	97.7	95.8	95.6	96.4
[10]	94.2	92	91.4	92.5	86.5	87.1	86.8	86.8	96.9	94.5	94.6	95.3
[32]	90.4	88.7	87.2	88.8	88.5	88.6	88.4	88.5	95.8	93.3	93.4	94.2
[33]	96.0	93.6	93.0	94.2	90.5	90.7	90.7	90.6	97.9	96.1	95.9	96.6
[34]	91.6	89.3	88.8	89.9	87.9	88.2	88.2	88.1	96.8	93.3	93.6	94.6

We have followed the proposed pipeline for all the models, where two privacy-protecting modalities are fused using the fusion algorithm and thereafter an uncovered RGB image is generated by the GAN architecture. The performance of the proposed method and SOTA models for estimation of in-bed human pose of the generated image is tabulated in Table VII. PCKh metric at 0.5 is used to evaluate all the SOTA models for different fused modalities and under different cover conditions (c0=no cover, c1=thin cover, c2=thick cover). The experimental results show that the fusion of depth and LWIR outperforms the other two pairs of fused modalities (LWIR+PM) and (PM+depth) for all the cover conditions by all the SOTA models and the proposed pose estimation method. This experiment clearly shows the supremacy of the fused pair (depth+LWIR) over the other two pairs. The experiment not only marks the highest score for the said pair of modalities but also achieves the highest score of PCKh@0.5 by the proposed pose estimation method which is a modified version of HRNet. Both depth and LWIR are not only privacy-protecting modalities but also inexpensive and easily accessible as compared to pressure maps which fail to achieve significant scores even after fusion with either of the modalities. The performance of pose estimation under no cover condition is inevitably higher, and the score decreases as the thickness of the cover increases. Features of neighboring scales are merged by the network architecture in [31] and [33]. A couple stacked UNet structure is used by [32], where the features are globally reused thereby making the network lightweight. [34] uses a CNN based multi-context attention mechanism for pose estimation.

TABLE VIII
COMPARISON OF PERFORMANCE FOR POSE ESTIMATION AT PCKH@0.5 BY PROPOSED METHOD AND RECENT FUSION BASED MODELS

Method	Year	PCKh@0.5
Proposed method	2023	**97.8**
[35]	2023	96
[36]	2022	95.8
[37]	2022	76.13

The proposed method seems to outperform on PCKh metric @ 0.5 when compared with recently published fusion-based models which are employed for in-bed pose estimation on the SLP dataset. Table VIII compares the proposed method with SOTA fusion-based pose estimation algorithms. [35] employs feature fusion by deep learning-based end-to-end fully trainable approach. Feature fusion-based approach is used by [36], where features of missing visible images are reconstructed. [37] have proposed unimodal pose estimation techniques. The experimental results show that the fusion of modalities provided better results for pose estimation.

Table IX shows the comparison of the proposed pipeline with [11] and [10] to compare the standard deviation and mean difference of predicted joints from the ground truth results. The low value for standard deviation and mean shows the supremacy of the model. The proposed method shows better results for both standard deviation and mean, this is because of the addition of feature fusion technique into the pipeline along with the synthetic image generation model which enhances the performance of pose estimation when compared with SOTA. Qualitative results for the proposed pipeline are shown in Fig. 8 where two modalities are considered as input and thereafter, they are pre-processed individually, and their features are fused. Henceforth, the GAN-based model generates uncovered RGB images and the pose is detected.

TABLE IX
COMPARISON OF STANDARD DEVIATION AND MEAN FOR POSE ESTIMATION BY PROPOSED METHOD AND SOTA

Model	STD	Mean
Proposed method	**0.086**	**0.142**
[11]	0.094	0.149
[10]	0.099	0.157

IV. CONCLUSION

In this research, a new IoMT approach has been developed for in-bed pose estimation exploiting deep learning techniques. The proposed model has been tested using a public dataset, SLP, for performance evaluation, and considerable improvements have been observed compared to the recent methods. The proposed method uses an AI-based generative model to recover images under thick cover situations. The deep learning-based pose estimation method has a direct impact on sleep applications and the potential to be integrated within smart healthcare systems using low-cost devices to monitor patients in the home and hospital environments. Moreover, the developed deep learning-based pose estimation approach can be exploited in other studies for real-time and full body pose estimation which has a big impact on smart healthcare as well as continuous monitoring of patients at home/hospital environments. Large-scale data collection at home/hospital environment can be done in future studies targeted at various healthcare applications such as for monitoring patients with pressure ulcers, sleep apnea disorder, acid reflux, elderly, neonatal, or pregnant women. The proposed architecture can also be used as a base to infer contact pressure from various modalities to eliminate the necessity of using pressure maps which will be another future direction of this research where augmented data using GAN can help improve the system performance.

REFERENCES

[1] A. M. Neill, S. M. Angus, D. Sajkov, and R. D. McEvoy, "Effects of sleep posture on upper airway stability in patients with obstructive sleep apnea," *American journal of respiratory and critical care medicine*, vol. 155, no. 1, pp. 199–204, 1997.

[2] D. M. Smith, "Pressure ulcers in the nursing home," *Annals of internal medicine*, vol. 123, no. 6, pp. 433–438, 1995.

[3] J. LaBuzetta, J. Hermiz, V. Gilja, and N. Karanjia, "Using accelerometers in the neurological icu to monitor unilaterally motor impaired patients," *Neurology*, vol. P3.204, 2016.

[4] J. Yuan, Z. Liu, and Y. Wu, "Discriminative video pattern search for efficient action detection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 33, no. 9, pp. 1728–1743, 2011.

[5] S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724–4732, 2016.

[6] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1281–1290, 2017.

[7] C. Ionescu, D. Papavaa, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[8] 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. 7, pp. 1–12, 2019.

[9] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Computer Vision and Image Understanding*, vol. 192, no. 102897, 2020.

[10] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," *In Proceedings of the European conference on computer vision (ECCV)*, pp. 466–481, 2018.

[11] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," 2019, pp. 5693–5703, In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

[12] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299, 2017.

[13] Z. Wang, M. A. Armin, S. Denman, L. Petersson, and D. Ahmedt-Aristizabal, "Video-based inpatient fall risk assessment: A case study," *In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pp. 2601—-2604, 2021.

[14] D. Ahmedt-Aristizabal, S. Denman, K. Nguyen, S. Sridharan, S. Dionisio, and C. Fookes, "Understanding patients' behavior: Vision-based analysis of seizure disorders," *IEEE journal of biomedical and health informatics*, vol. 23, no. 6, pp. 2583–2591, 2019.

[15] S. Liu, X. Huang, F. Xiaofei, N. Fu, C. Li, Z. Su, and S. Ostadabbas, "Simultaneously-collected multimodal lying pose dataset: Enabling in-bed human pose monitoring," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 1106–1118, 2023.

[16] H. M. Clever, P. L. Grady, G. Turk, and C. C. Kemp, "Bodypressure - inferring body pose and contact pressure from a depth image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 137–153, 2023.

[17] S. Liu and S. Ostadabbas, "Pressure eye: In-bed contact pressure estimation via contact-less imaging," *Medical Image Analysis*, vol. 87, pp. 1361–8415, 2023.

[18] C. V. Kamath, S. Liu, and S. Ostadabbas, "Privacy-preserving in-bed pose and posture tracking on edge," *44th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC).*, pp. 3365–3369, 2022.

[19] P. Climent-Pérez and F. Florez-Revuelta, "Protection of visual privacy in videos acquired with rgb cameras for active and assisted living applications," *Multimed Tools Appl*, vol. 80, pp. 23649—-23664, 2021.

[20] D. G. Kyrollos, A. Fuller, K. Greenwood, J. Harrold, and J. R. Green, "Under the cover infant pose estimation using multimodal data," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, Eds. Curran Associates, Inc.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. 2015, vol. 9351, pp. 234–241, Springer, LNCS.

[23] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," in *Lecture Notes in Computer Science, Computer Vision – ECCV*. 2016, vol. 9908, pp. 318—-335, Springer, Cham.

[24] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2017, p. 936–944, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[26] Y. Yu, W. Zhang, and Y. Deng, *Frechet Inception Distance (FID) for Evaluating GANs*, China University of Mining Technology Beijing Graduate School, 2021.

[27] F. A. Fardo, V. H. Conforto, F. C. de Oliveira, and P. S. Rodrigues, "A formal evaluation of psnr as quality measurement parameter for image segmentation algorithms," 2016, p. 1123–1134, arXiv preprint arXiv:1605.07116.

[28] J. Nilsson and T. Akenine-Möller, "Understanding ssim.," 2020, p. 1123–1134, arXiv preprint arXiv:2006.13846.

[29] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2017, p. 1123–1134, IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[30] T. C. Wang, M. Y. Liu, J. Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807, 2018.

[31] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," 2016, pp. 483—-499, Proc. Eur. Conf. Comput. Vis.

[32] G. Gkioxari, A. Toshev, and N. Jaitly, "Chained predictions using convolutional neural networks," 2016, pp. 728—-743, Proc. Eur. Conf. Comput. Vis.

[33] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," 2017, pp. 1281—1290, Proc. IEEE Int. Conf. Comput. Vis.,.

[34] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," 2017, pp. 1831—1840, Proc. IEEE Conf. Comput. Vis. Pattern Recognit.

[35] T. Dayarathna, T. Muthukumarana, R. Yasiru, C. de Silva S. Denman, A. Pemasiri, and D. Ahmedt-Aristizabal, "Privacy-preserving in-bed pose monitoring: A fusion and reconstruction study," *Expert Systems with Applications*, vol. 213, pp. 119–139, 2023.

[36] T. Cao, M. A. Armin, S. Denman, L. Petersson, and D. Ahmedt-Aristizabal, "In-bed human pose estimation from unseen and privacy-preserving image domains," 2022, pp. 1—5, IEEE 19th International symposium on biomedical imaging.

[37] M. Afham, U. Haputhanthri, J. Pradeepkumar, M. Anandakumar, A. De Silva, and C. U. S. Edussooriya, "Towards accurate cross-domain in-bed human pose estimation.," 2022, p. 2664–2668, IEEE International conference on acoustics, speech and signal processing (ICASSP).