Research progress briefing

Lightweight Dual-attention Hourglass Network for RGB-D In-bed Pose Estimation

Presenter: 林保羅 (Paulo Linares)

指導教授: 傅楸善(Chiou-Shann Fuh) 博士

E-mail: d12922028@ntu.edu.tw

September 16, 2025

Problem

Can you spot the body joints on these images?

(a) ***(b)***

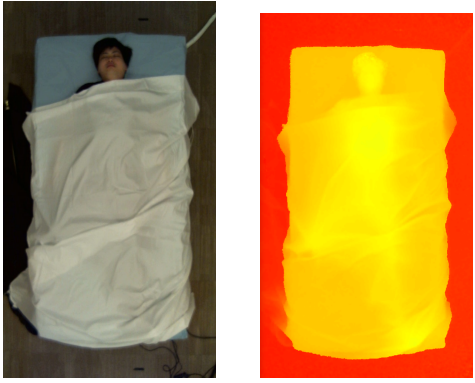
Figure 1: *RGB images of an in-bed patient.*

Problem (cont.)

- This kind of images can be obtained at a hospital room with a regular RGB camera.
- It is important to **monitor in-bed patients** at a hospital in **real-time**.
- Finding the body joint locations of a patient would allow to analyze his behavior (action recognition) so that nurses and doctors can take care of him on time.
- **Blanket occlusion** (usually present in images taken from patients laying on a hospital bed) leads to an inaccurate body joint localization. Thus, an action recognition model using just RGB images (regular color images) might not be the best option.
- In computer vision, the shape or **pose of a person** in an image is denoted as a collection of key points representing specific joints connecting the main parts (e.g. limbs, head, torso) of a human body. The goal of **Human Pose Estimation (HuPE)** is to estimate or predict the location of these key points in a 2D or 3D environment .

Problem (cont.)

Can you spot the body joints by combining the info in these images?



(a) ***(b)***

Figure 2: *RGB-D images of an in-bed patient: (a) RGB part; (b) Depth part.*

Problem (cont.)

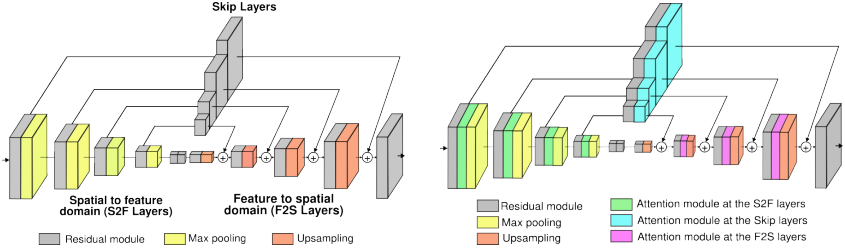
Human Pose Estimation vs. In-bed Pose Estimation

Task	Main goal	Scenario	Production-scale models/toolkits
Human Pose Estimation	Localize a person's body joints given an image in any modality (RGB, depth, IR, PM)	▶ The person's body is mostly visible. ▶ The person is performing any kind of activity at any environment.	▶ Google MediaPipe ▶ MoveNet ▶ OpenPose ▶ MMPose ▶ Ultralytics YOLO-pose
In-bed Pose Estimation		▶ The person's body is likely to be almost completely occluded by a blanket. ▶ The person is lying down on a hospital bed. Its actions are limited to small movements within the bed (rolling, sitting).	**Quite a few** (Only pre-trained models for academical purposes)

PM: Pressure Maps, IR: Infra-Red, RGB: Red, Green, Blue, YOLO: You-Only-Look-Once.

Method

- Employ RGB-D images. An RGB-D image has 4 channels (3 color channels + 1 depth channel).
- The depth channel captures the distance between the camera and the object (in meters).
- A shallow DNN-based model is proposed. Aiming to achieve **real-time processing during patient monitoring in edge devices**. The **number of parameters is kept low**, whereas the **accuracy should be acceptable**.
- The proposed model is based on Stacked Hourglass Networks (A. Newell, 2016) [1].



(a) ***(b)***

Figure 3: Proposed Hourglass module including attention mechanisms: *(a) Original Hourglass module; (b) Attention mechanisms incorporated into the ResNet modules comprising an Hourglass.*

RGB-D: Red, Green, Blue - Depth, DNN: Deep Neural Networks

Method (cont.)

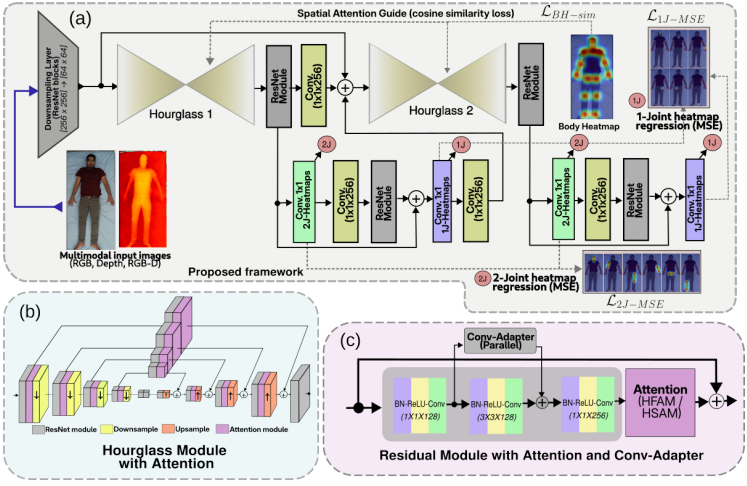


Figure 4: Proposed Dual-Attention Hourglass Network architecture: *(a) Overall framework of the proposed Hourglass network with attention modules and supervision on body joints and limbs; (b) Incorporation of attention modules to the layers of an Hourglass; (c) Attention module incorporated into a ResNet module with a Conv-Adapter.*

Method (cont.)

Incorporation of attention into a ResNet module

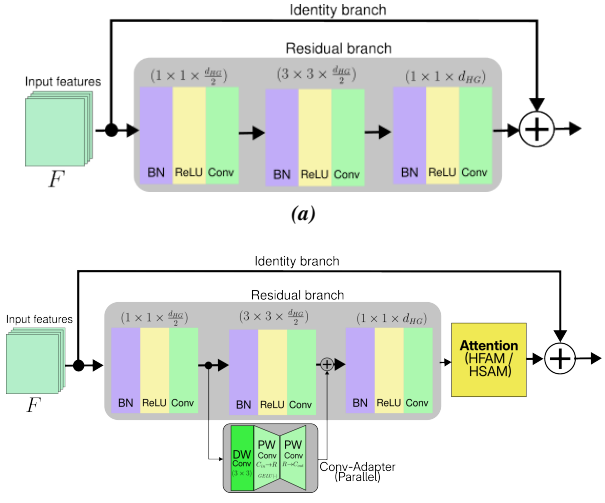


Figure 5: Incorporation of the proposed attention mechanisms into a Residual module: *(a) Original bottleneck ResNet module with pre-activation; (b) Incorporation of the mentioned attention modules and Conv-Adapter[2] into a ResNet module. The Conv-Adapter and attention module are not included during pre-training*

Method (cont.)

Main Contributions

1. **Customized Attention Mechanisms (AM)**: Two different AMs are added to the residual modules. One AM is designed for the S2F part of the HG. Another AM is proposed for the F2S part.
2. **Multi-joint ground-truth heatmap generation**: Introduction of more heatmaps (2-joint heatmaps) to increase the level of supervision.
3. **Stage-specific heatmaps**: The heatmaps are generated from the ground truth pose according to the hourglass position (stage number) within the model.
4. **MSE-based loss with attention guide**: Combines MSE for heatmaps (1-joint, 2-joint) regression, and *cosine similarity* ($\mathcal{L}_{\text{BH-sim}}$) for guiding the generation of spatial attention maps.

MSE: Mean Squared Error

Method (cont.)

Proposed attention mechanisms

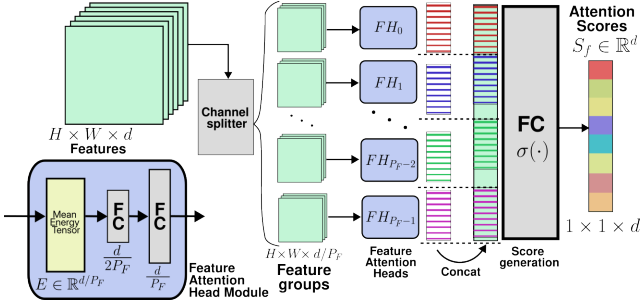
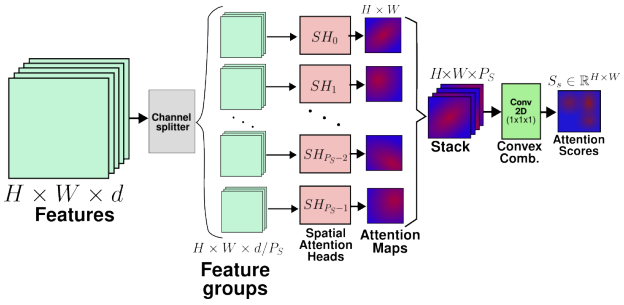


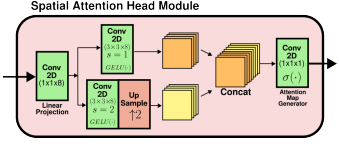
Figure 6: *Proposed Feature Attention Mechanism (HGFAM).*

FC: Fully connected layer

Method (cont.)



(a)



(b)

***Figure 7: Proposed Spatial Attention Mechanism (HGSAM)**: (a) Overall architecture of the HGSAM module comprising P_S spatial attention heads; (b) Proposed spatial attention head used to generate an attention map.*

Method (cont.)

Table 1: *Summarization of the hyperparameters involved in HGFAM and HGSAM*

Module	Parameter	Description	Value
HGFAM	P_F	Number of feature attention heads	4
	dF_{hidden}	Head's hidden layer dimensionality	$\frac{d}{2P_F}$
	dF_{head}	Head's last FC layer output dimensionality	$\frac{d}{P_F}$
	dF_{sgen}	HGFAM's score generator dimensionality	d
HGSAM	P_S	Number of spatial attention heads	4
	dS_{proj}	Output channel number at the head's linear projection	8
	dS_{extra}	Channel number for the 2 (3x3) Conv layers at the head.	8
	dS_{head}	Head's output channel number	1
	dS_{sgen}	Number of channel at the HGSAM's score generator	1

Method (cont.)

Ground-truth Heatmaps generation

Proposed Heatmap generation scheme (1-joint , 2-joint heatmaps):

- Conventionally, the ground truth heatmaps for training HuPE models are generated in a one-joint per heatmap basis.
- A 1-joint heatmap is represented by a 2D Gaussian function in a discrete space. Where the position of the k^{th} joint is adopted as the mean of the Gaussian (σ is kept fixed in previous works). The main problem is that this kind of heatmap is mostly sparse.
- This work introduces the concept of 2-joint heatmaps. The goal of including these heatmaps is to reduce the sparsity while conveying information related to body limbs (e.g. forearms, legs).

Method (cont.)

1-Joint Heatmaps

$$H_k(\vec{x}|\vec{x_k}, \sigma) = \exp\left(-\frac{1}{2}\frac{\|\vec{x} - \vec{x_k}\|^2}{\sigma^2}\right) \tag{1}$$

$\vec{x_k}$: 2D location of the k^{th} joint; σ: standard deviation (fixed, dynamic according to the HG stage).

2-Joint Heatmaps

$$D_{[x,k_0,k_1]} = \|\vec{x} - \vec{x_{k_0}}\| + \|\vec{x} - \vec{x_{k_1}}\| - \|\vec{x_{k_0}} - \vec{x_{k_1}}\| \tag{2}$$

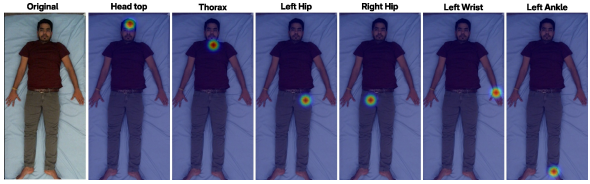
$$\mathcal{H}_{[k_0,k_1]}(\vec{x}|\sigma) = \frac{1}{2}\left[H_k(\vec{x}|\vec{x_{k_0}}, \sigma) + H_k(\vec{x}|\vec{x_{k_1}}, \sigma)\right] \tag{3}$$

$$H_{[k_0,k_1]}(\vec{x}|\sigma) = \mathcal{K}_{1J}\mathcal{H}_{[k_0,k_1]}(\vec{x}|, 1.5\sigma) + \mathcal{K}_{2J}\exp\left(-\frac{1}{2}\frac{D_{[x,k_0,k_1]}}{\sigma}\right) \tag{4}$$

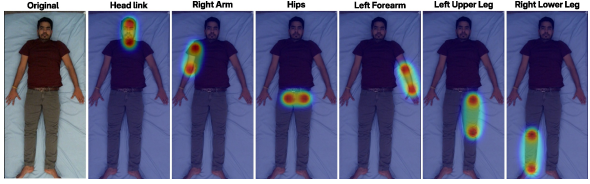
$$\mathcal{K}_{1J}, \mathcal{K}_{2J} > 0; \quad \mathcal{K}_{1J} + \mathcal{K}_{2J} = 1.0$$

$\vec{x_{k_0}}, \vec{x_{k_1}}$: 2D locations of the joints comprising a body part.

Method (cont.)



(a)



(b)

Figure 8: Proposed ground truth heatmap generation scheme: *(a) Sample 1-joint heatmaps (conventional body joint heatmaps); (b) Samples of the proposed 2-joint heatmaps.*

Method (cont.)

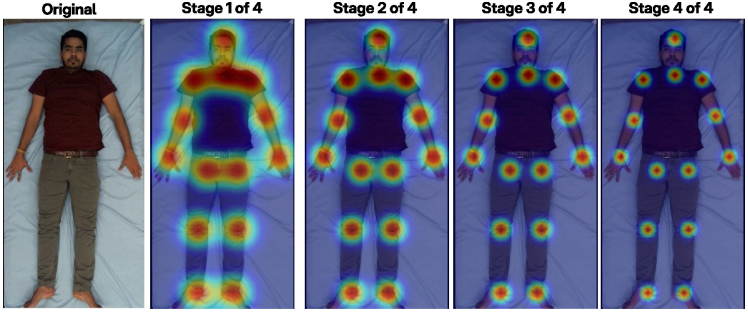


Figure 9: *Proposed stage-specific ground-truth heatmaps.A sample from the SLP dataset was employed. Heatmaps are generated according to (1) with* $\sigma = 1.3$.

$$\sigma_n = \gamma^{N-n}\sigma_N; \quad \gamma \geq 1.0 \tag{5}$$

SLP: Simultaneously-collected multimodal Lying Pose Dataset

Method (cont.)

Interconnection between HGNets

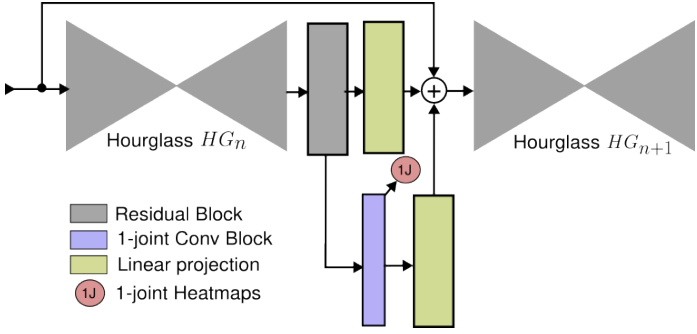


Figure 10: *Interconnection between two consecutive Hourglass Networks (Original version)*

Method (cont.)

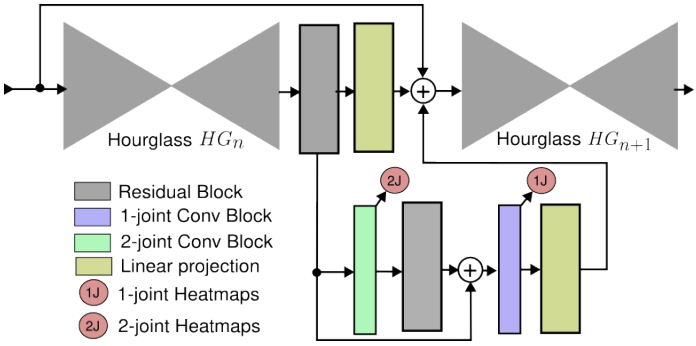


Figure 11: *Interconnection between two consecutive Hourglass Networks (Proposed version regarding 2-joint HMs)*

Method (cont.)
Experimental trials

1. Ablation studies regarding the architecture.
   - Stacked HGNets with no attention mechanisms.
   - Utilization of feature attention mechanisms (HGFAM).
   - Utilization of spatial attention mechanisms (HGSAM).
   - Stacked HGNets + HGFAM + HGSAM
2. Ablation studies regarding the image modality.
   - Depth (just 1 channel).
   - RGB-D (3 color channels + 1 depth channel).
3. Studies on the heatmap generation parameters.
   - Fixed vs. Variable standard deviation.
   - 1-joint heatmaps vs. 1-joint+2-joint heatmaps.
4. Qualitative evaluation.

Method (cont.)

Model ID	Attention Mechanisms			Heatmaps	
	S2F layers	F2S layers	Skip layers	1J	2J
HG-1	✗	✗	✗	✓	✗
HG-2	✗	✗	✗	✓	✓
HG-FNN-1	HGFAM	✗	✗	✓	✗
HG-NSN-1	✗	HGSAM	✗	✓	✗
HG-FNN-2	HGFAM	✗	✗	✓	✓
HG-FFN-1	HGFAM	HGFAM	✗	✓	✗
HG-FSN-1	HGFAM	HGSAM	✗	✓	✗
HG-SFN-2	HGSAM	HGFAM	✗	✓	✓
HG-FSN-2	HGFAM	HGSAM	✗	✓	✓
HG-FSF-2	HGFAM	HGSAM	HGFAM	✓	✓

The model IDs follow the structure:
HG-XYZ-N. X: attention used at S2F layers, Y: attention used at F2S layers, Z: attention used at Skip layers, N: type of heatmaps (1 or 2 joint heatmaps). 1J: 1-Joint, 2J: 2-Joint, S: Spatial Attention, F: Feature Attention, N: No Attention.

Table 2: *Summarization of the models employed during the ablation studies*

Method (cont.)

Implementation and Training details

Implementation

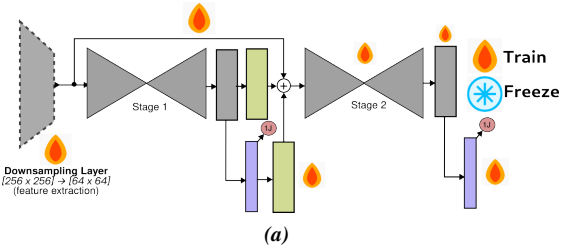
- **Python version:** 3.11
- **Framework and backend:** Keras 3.8, Tensorflow 2.18
- **Image modality support:** Depth, RGB-D
- **Number of features at the Hourglasses**: $d_{HG} = 256$.
- **Spatial dimensions of heatmaps and features**: 64×64.

Training

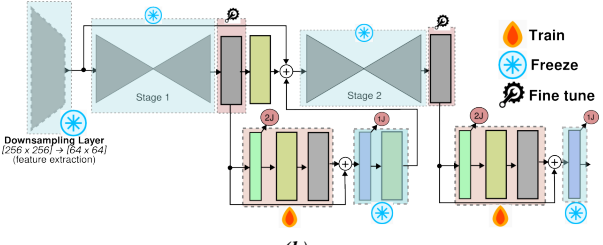
- **Optimizer:** Adam
- **Loss:** $\beta_0 \mathcal{L}_{1J-MSE} + \beta_1 \mathcal{L}_{2J-MSE} + \beta_2 \mathcal{L}_{BH-SIM}$
- **Scheduler:** Reduce LR on plateau (customized, reduce by 0.4 with tolerance of 4 epochs)
- Training in **4 stages:**
  1. Pre-training the baseline with 1-joint heatmaps only ($LR = 2.5 \times 10^{-4}$)
  2. Freeze the core hourglasses, add 2-joint layers, train for up to 20 epochs. ($LR = 2.5 \times 10^{-4}$).
  3. Incorporate attention moduless into the Hourglass modules and fine-tune Conv-Adapter within Hourglasses for up to 30 epochs ($LR = 10^{-5}$)
  4. Fine tune the entire model for up to 20 epochs. ($LR = 10^{-6}$)

Nvidia RTX 3090: 12.5 Hours baseline pre-training, Inference: 25 Frames Per Second
LR: Learning Rate

Method (cont.)



(a)



(b)

Figure 12: *Example depicting the training during Stage 1 and Stage 2: (a) Pre-training the baseline HG-1 (no attention mechanisms, just 1-Joints) ; (b) Training the blocks related to 2-Joints during the second stage.*

Method (cont.)

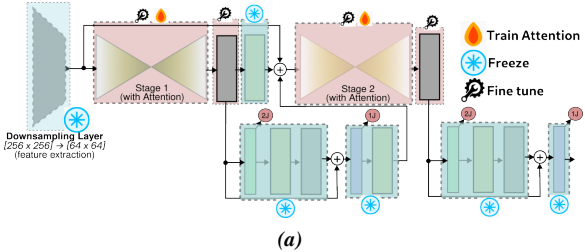


Figure 13: *Example depicting the training during Stages 3 and 4: (a) Inclusion of attention modules within ResNet modules (see Fig. 8b); (b) Final training stage (fine-tuning).*

Experimental Results

Selected benchmarks

- **Simultaneously-collected multimodal Lying Pose Dataset (SLP) [3]:** Aimed for in-bed pose estimation RGB images, Depth maps, LWIR images, Pressure Mat Maps. Different sensors are employed for each modality (i.e. image modalities are not aligned). Size: 30GB (†8GB)
- **Multi-View Kinect Dataset (MKV) [4]:** Aimed for RGB-D pose estimation and Robotic Task Learning. The RGB and depth frames are aligned before conducting experimental trials. Size: 60GB
- **UTD-MHAD [5]:** Aimed for action recognition. Comprises RGB Videos + Depth sequences. RGB and Depth sequences have not the same dimensions neither the same number of samples. Only the depth frames are employed. Size: 1.4GB
- **DCCV-BedPose**: Aimed at testing in-bed pose estimation models. Comprises RGB-D images obtained with an Intel RealSense L515 camera and follows the structure of SLP. Data was collected from 4 subjects, with 20 samples per subject (3 covering scenarios).

UTD-MHAD: University of Texas, Dallas - Multimodal Human Action recognition Dataset
DCCV: Digital Camera and Computer Vision Laboratory

Experimental Results (cont.)

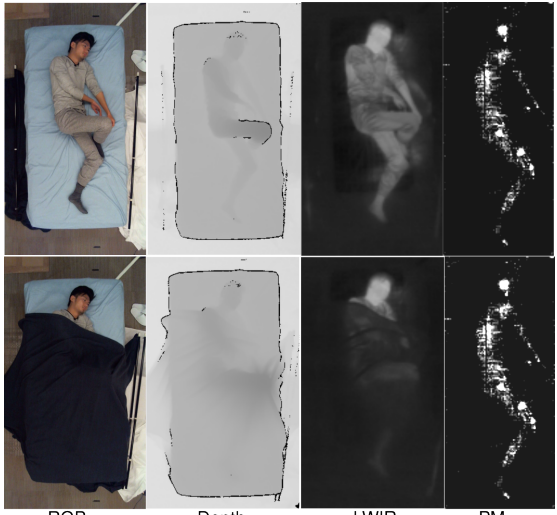
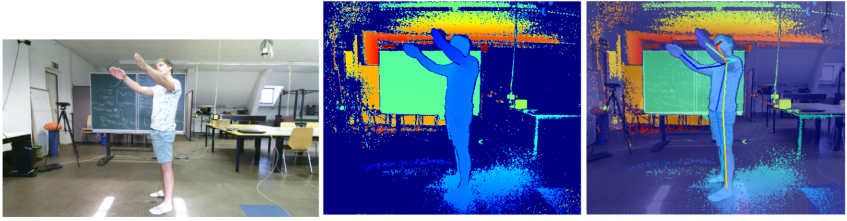


Figure 14: *Samples from SLP.*
The images were cropped and resized for visualization.

Experimental Results (cont.)

Frame Alignment on MKV

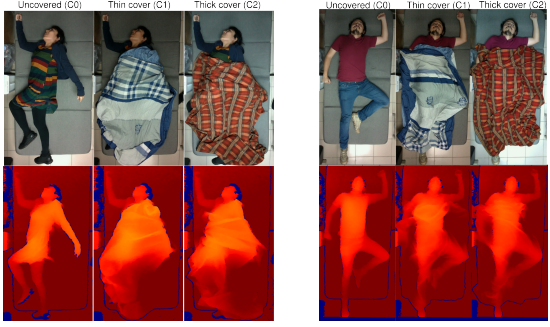


(a) *(b)* *(c)*

Figure 15: *Frame alignment on MKV: (a) Original RGB frame; (b) Original depth frame; (c) RGB and Depth blended after alignment.* *Frame alignment is conducted by applying a geometric transformation to the RGB image from a homography matrix.*

Experimental Results (cont.)

DCCV-BedPose



(a) ***(b)***

Figure 16: *Samples from DCCV-BedPose: (a) Samples of subject 1; (b) Samples of subject 2.*

Details:

- Aimed at in-bed pose estimation and action recognition.
- 3 cover scenarios.
- 20 samples per subject.
- So far only 3 subjects.
- Includes RGB-D sequences of subjects performing 12 actions.

Experimental Results (cont.)

Evaluation Metrics

- **Percentage of Correct Keypoints (PCKh@0.5):** A keypoint is regarded as a correct one if the distance between its predicted and real (ground truth) locations are within a threshold distance. This threshold distance is computed by multiplying the GT head length by a factor of 0.5.
- **Mean Per Joint Error (MPJE):** Average error distance (in pixels) between the predicted joint locations and the ground truth.
- **Latency and FPS**: The throughput of all the models is between 20-30 FPS (running on a AMD Ryzen 5 5600 CPU - without GPU).

Experimental Results

Results on SLP

(In-bed pose estimation)

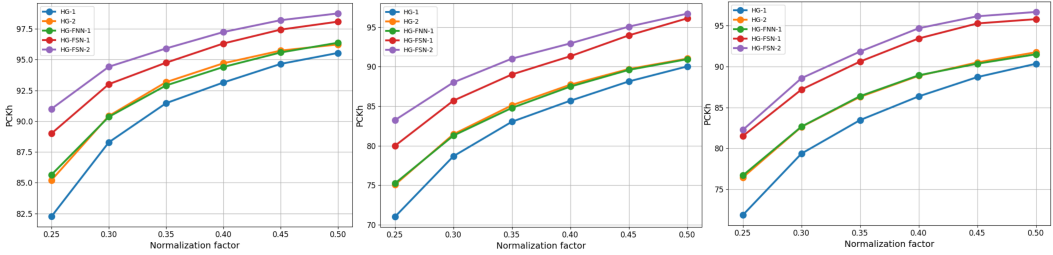
Experimental Results (cont.)

Table 3: *HuPE performance on SLP under blanket cover (with cover) and uncovered (no-cover) occlussion scenarios*

Method	SLP (no-cover)		SLP (with cover)	
	PCKh(↑)	MPJE(↓)	PCKh(↑)	MPJE(↓)
HG-1	95.54	1.265	90.19	1.874
HG-2	96.23	1.263	91.39	1.797
HG-FNN-1	96.37	1.223	91.23	1.758
HG-SNN-1	96.40	1.223	91.49	1.757
HG-FNN-2	96.88	1.218	91.64	1.758
HG-FSN-1	98.08	1.206	95.87	1.741
HG-SFN-1	**98.09**	1.215	94.92	1.751
HG-FSN-2	98.08	**1.198**	**96.16**	**1.729**

Depth is employed as the image modality.
All the models are trained with stage specific heatmaps ($\gamma = 1.35$).

Experimental Results (cont.)



(a) *(b)* *(c)*

Figure 17: Per-joint HuPE results of the proposed models on SLP: *(a) Without cover; (b) Thin cover; (c) Thick cover.*

Experimental Results (cont.)

Table 4: *Comparison with SOTA models on SLP according to HuPE performane (PCKh@0.5) and Model's size (number of parameters).* `C0: No-cover, C1: Thin cover, C2: Thick cover.`

Method	Modality	PCKh@0.5(↑)				Params
		C0	C1	C2	Overall	(↓)
SHG + DAug + KD [6]	LWIR	-	-	-	76.13	40.8M
HRNet + iAFF [7]	RGB + LWIR	96.5		92.5	94.3	60.0M
HRNet + Fusion [8]	Depth + LWIR	-	-	-	**97.3**	36.4M
SHG† [3]	Depth	97.6	96.1	95.8	96.5	12.6M
HG-1	Depth	95.54	90.04	90.34	91.98	**6.40M**
HG-2	Depth	96.23	91.03	91.75	93.00	6.43M
HG-FNN-1	Depth	96.37	91.70	91.01	93.05	6.9M
HG-FNN-1	RGB-D	97.02	90.85	90.50	92.94	6.9M
HG-FSN-1	Depth	98.08	96.12	95.88	96.65	7.03M
HG-FSN-2	Depth	**98.08**	**96.71**	**96.08**	96.88	7.09M

† 2 ResNet blocks per ResNet module are employed. Our proposed model only uses 1 ResNet block per module.

SHG: Stacked Hour Glass; iAFF: iterative Attentional Feature Fusion; KD: Knowledge Distillation; DAug: Data Augmentation; HRNet: High-Resolution Network.

Experimental Results (cont.)

Why LWIR (thermal) images may not be suitable for monitoring an in-bed patient ?

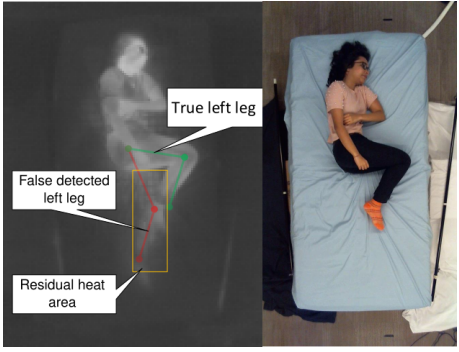


Figure 18: *Ghosting effect present in LWIR (thermal) images captured in an in-bed scenario.*

"... as human moves in the bed, the "heat residue" of the previous pose will result in ghost temperature patterns as the heated area needs time to gradually diffuse heat. [3]"

Experimental Results (cont.)

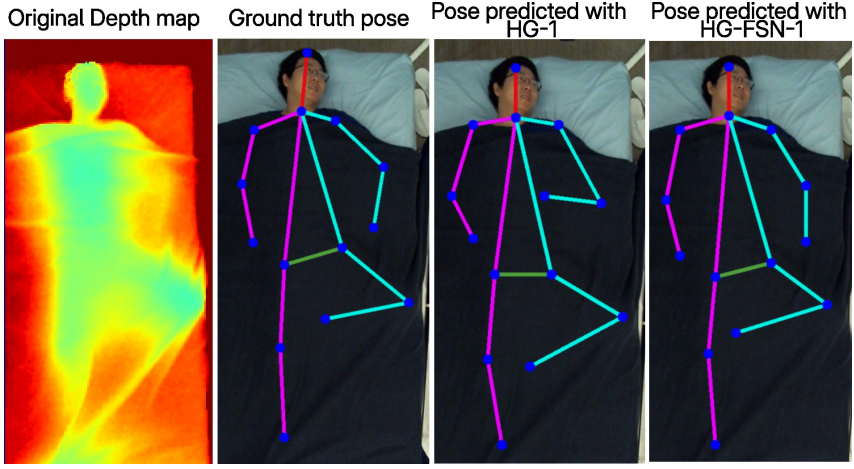
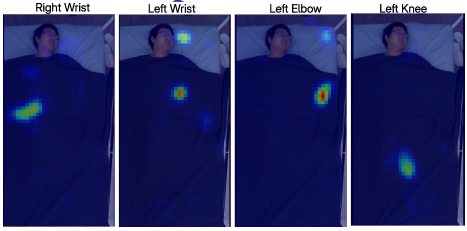
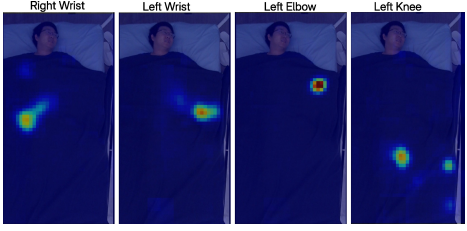


Figure 19: *Qualitative results (heatmaps and pose) on an image from SLP without blanket occlussion.*

Experimental Results (cont.)



(a)



(b)

Figure 20: *Qualitative results (heatmaps and pose) on an image from SLP with blanket occlussion.*

Experimental Results (cont.)

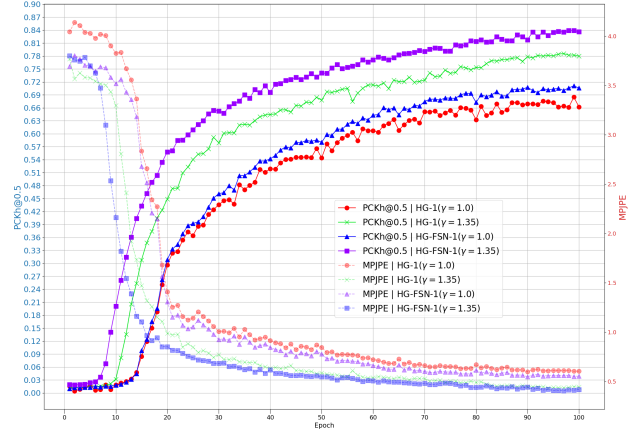


Figure 21: Validation metrics (PCKh@0.5, MPJPE) monitored along the training of the models with fixed and stage-specific ground truth heatmaps.

- Smaller variants of the models are trained using both, heatmaps with a fixed standard deviation ($\gamma = 1.0$), and a variable standard deviation ($\gamma = 1.35$).
- The model size is reduced by decreasing the number of features from 256 to 64.
- The influence of γ during training is amplified by increasing the number of stages from 2 to 5.
- Model size: around $1.0M$ parameters.

Experimental Results

Results on MKV

(Regular pose estimation)

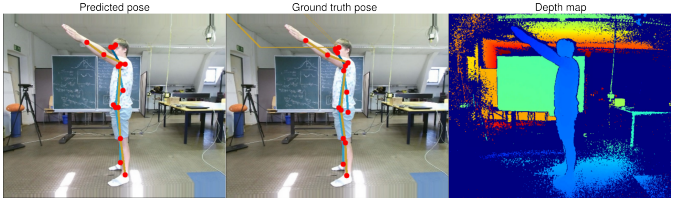
Experimental Results (cont.)

Table 5: *HuPE performance of the proposed models on MKV*

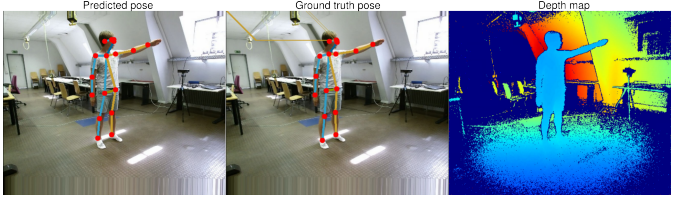
Method	d_{HG}	# stages	Modality	PCKh@0.5(↑)				
				View 1	View 2	View 3	View 4	Overall
HG-1	256	2	Depth	90.02	89.15	92.12	89.92	89.35
HG-FNN-1	256	2	Depth	91.23	91.70	92.51	91.54	90.78
HG-FSN-1	256	2	Depth	92.44	92.45	92.55	92.08	91.41
HG-FSN-2	256	2	Depth	93.51	94.05	94.66	93.16	92.86
HG-FSN-2	256	2	RGB-D	**93.77**	**94.87**	**95.49**	**93.50**	**93.42**
HG-FSN-2	64	4	RGB-D	88.65	87.56	90.15	89.05	88.85

HuPE: Human Pose Estimation.
d_{HG}: Number of features used at the Hourglass modules.

Experimental Results (cont.)



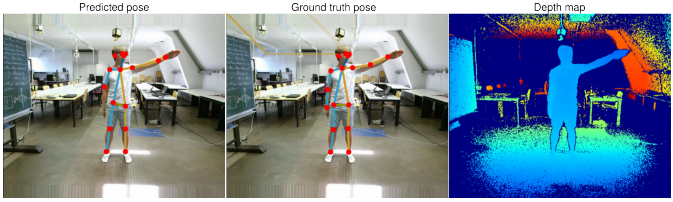
(a)



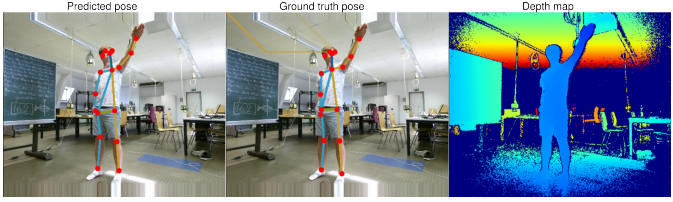
(b)

Figure 22: *Qualitative results on MKV: (a) Camera view 1; (b) Camera view 2.*

Experimental Results (cont.)



(a)



(b)

Figure 23: *Qualitative results on MKV: (a) Camera view 3; (b) Camera view 4.*

Experimental Results

Results on UTD-MHAD

(Regular pose estimation)

Experimental Results (cont.)

Table 6: *Per-joint HuPE performance on UTD-MHAD regarding PCKh@0.5*

Model	Ank./ Feet	Knee	Hip	Elb.	Sho.	Wri./ Hand	Head	Total
HG-1	94.51	96.21	98.37	97.08	98.03	95.77	98.06	95.83
HG-2	94.52	97.00	98.40	97.91	99.01	95.78	98.14	96.18
HG-FNN-1	94.43	97.01	98.34	97.85	99.12	95.03	98.06	96.04
HG-FSN-1	95.00	97.14	98.36	97.93	**99.12**	96.28	98.22	96.36
HG-FSN-2	**95.21**	**97.20**	**98.36**	**98.02**	99.11	**96.48**	**98.23**	**96.44**

Experimental Results

Results on DCCV-BedPose

The models trained on SLP are employed.

(No samples from this dataset were used for training.)

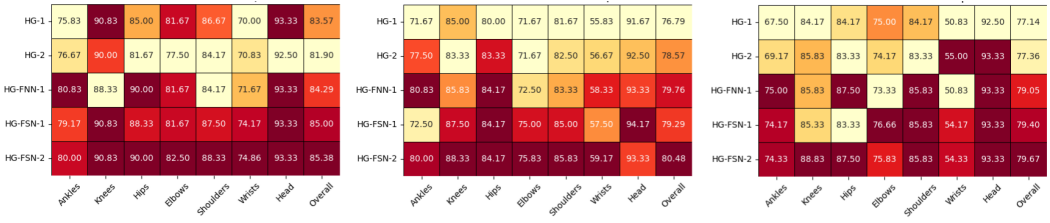
Experimental Results (cont.)

Table 7: *HuPE performance on DCCV-BedPose according to PCKh@0.5*

Model	Modality	PCKh@0.5(↑)			
		C0	C1	C2	Overall
HG-1	Depth	83.5	76.79	77.14	79.14
HG-2	Depth	81.90	78.57	77.36	79.28
HG-FNN-1	Depth	84.29	79.76	79.05	81.03
HG-FSN-1	Depth	85.00	79.29	79.40	81.23
HG-FSN-1	RGB-D	**87.02**	78.81	78.76	81.53
HG-FSN-2	Depth	85.38	**80.48**	**79.67**	**81.84**

C0: No occlusion, C1: Thin occlusion, C2: Thick occlusion

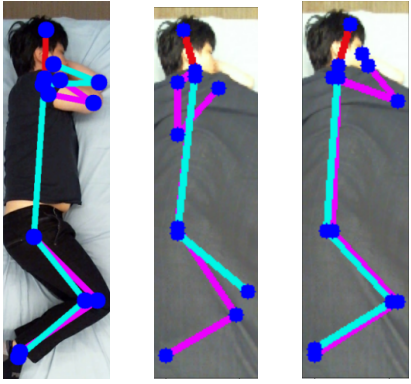
Experimental Results (cont.)



(a) ***(b)*** ***(c)***

Figure 24: Per-joint HuPE results on DCCV-BedPose*: (a) Without cover; (b) Thin cover; (c) Thick cover.*

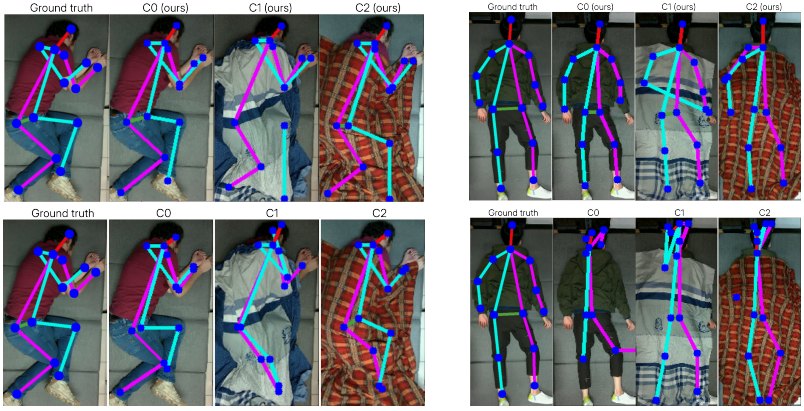
Experimental Results (cont.)



(a) ***(b)*** ***(c)***

Figure 25: Qualitative comparison on pose estimation between the baseline HG-1 and HG-FSN-2*: (a) Ground truth pose; (b) Pose predicted with the baseline HG-1; (c) Pose predicted with HG-FSN-2.*

Experimental Results (cont.)



(a) ***(b)***

Figure 26: Qualitative comparison on DCCV-BedPose between our method and [3]: *(a) Sample depicting a resting position on the side; (b) Sample depicting a prone position. Top: Results obtained with our method (HG-FSN-2). Bottom: Results obtained with [3].*

Conclusion

- The proposed models and attention mechanisms have been implemented.
- The proposed ground-truth generation scheme in conjuction with the multi-stage training regime improve the performance of the original Stacked Hourglass Network during pose estimation.
- All our models comprise 2 stacked Hourglass Nets, with a total of $6.8M$ - $7.1M$ parameters (12M parameters for [3]). **Model Size: 48-55 Megabytes.** (150 MBs for [3])
- The principal metric used for assessing the model is the PCK (Percentage of Correct Keypoints). This metric measures how many keypoints are localized correctly within a specific distance from the ground-truth locations. The threshold distance is computed based on the distance between the head's top and neck (PCKh@0.5).
- So far the model has been tested on SLP(*Simultaneously-collected multimodal Lying Pose*) and DCCV-BedPose for in-bed pose estimation. The proposed model has attained **96.8%** of **PCKh@0.5** on the validation set of SLP.
- The experimental results on MKV and UTD-MHAD are employed just to verify the performance boost obtained from the attention mechanisms.

Summary

- Methodology: Use a modified version of the Stacked Hourglass Network to localize the body joints of an in-bed patient by using Depth or RGB-D images.
- Steps:
  1. RGB-D image preprocessing: Noise removal in the depth channel (e.g. holes), ensure coherence between the color channels and depth channel (in case the images are obtained with different cameras).
  2. Generate a dataset comprising coherent RGB-D images, and their corresponding ground truth body joint locations.
  3. Generate the ground truth heatmaps from the body joint locations.
  4. Train the baseline Stacked Hourglass Network with Depth / RGB-D images, and the proposed heatmap generation scheme.
  5. Follow the proposed multi-stage training regime to incorporate aditional modules (attention mechanisms, 2J heatmap blocks) to the baseline.
  6. Evaluate the performance of the human pose estimation model using the PCK (Percentage of Correct Keypoints) metric.
- Results: Real-time body joints localization from Depth / RGB-D images (containing one person) even under blanket occlusion.

References

[1] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision (ECCV)*, Springer, 2016.

[2] H. Chen, R. Tao, H. Zhang, Y. Wang, X. Li, W. Ye, J. Wang, G. Hu, and M. Savvides, "Conv-Adapter: Exploring parameter efficient transfer learning for convnets," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1551–1561, 2024.

[3] S. Liu, X. Huang, 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.

[4] C. D. W. B. Christian Zimmermann, Tim Welschehold and T. Brox, "3d human pose estimation in RGBD images for robotic task learning," in *IEEE International Conference on Robotics and Automation, ICRA*, 2018.

[5] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 168–172, 2015.

[6] 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," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2664–2668, 2022.

[7] 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," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2022.

[8] T. Dayarathna, T. Muthukumarana, Y. Rathnayaka, S. Denman, C. de Silva, A. Pemasiri, and D. Ahmedt-Aristizabal, "Privacy-preserving in-bed pose monitoring: A fusion and reconstruction study," *Expert Systems with Applications*, vol. 213, p. 119139, 2023.