Robust Multi-modal 3D Patient Body Modeling*

Fan Yang[1,2][0000−0003−1535−447X], Ren Li[1][0000−0003−2998−7104], Georgios Georgakis[1,3][0000−0002−5501−2311], Srikrishna Karanam[1][0000−0002−7627−7765], Terrence Chen[1], Haibin Ling[4][0000−0003−4094−8413], and Ziyan Wu[1][0000−0002−9774−7770]

[1] United Imaging Intelligence, Cambridge MA, USA
[2] Temple University, Philadelphia PA, USA
[3] George Mason University, Fairfax VA, USA
[4] Stony Brook University, Stony Brook NY, USA
{first.last}@united-imaging.com hling@cs.stonybrook.edu

Abstract. This paper considers the problem of 3D patient body modeling. Such a 3D model provides valuable information for improving patient care, streamlining clinical workflow, automated parameter optimization for medical devices *etc.* With the popularity of 3D optical sensors and the rise of deep learning, this problem has seen much recent development. However, existing art is mostly constrained by requiring specific types of sensors as well as limited data and labels, making them inflexible to be ubiquitously used across various clinical applications. To address these issues, we present a novel robust dynamic fusion technique that facilitates flexible multi-modal inference, resulting in accurate 3D body modeling even when the input sensor modality is only a subset of the training modalities. This leads to a more scalable and generic framework that does not require repeated application-specific data collection and model retraining, hence achieving an important flexibility towards developing cost-effective clinically-deployable machine learning models. We evaluate our method on several patient positioning datasets and demonstrate its efficacy compared to competing methods, even showing robustness in challenging patient-under-the-cover clinical scenarios.

Keywords: 3D patient pose and shape · multi-modal

1 Introduction

We consider the problem of 3D patient body modeling. Given an image of a patient, the aim is to estimate the pose and shape parameters of a 3D mesh that digitally models the patient body. Such a 3D representation can help augment existing capabilities in several applications. For instance, for CT isocentering, the 3D mesh can provide an accurate estimate of thickness for automated patient

* Fan Yang and Ren Li are joint first authors. This work was done during the internships of Fan Yang, Ren Li, and Georgios Georgakis with United Imaging Intelligence. Corresponding author: Srikrishna Karanam.

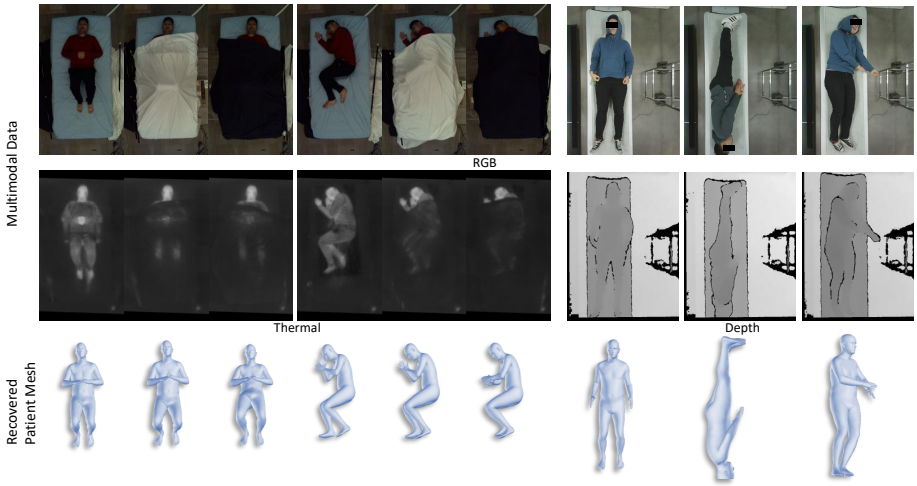

Fig. 1. We present a new approach for 3D patient body modeling that facilitates mesh inference even when the input data is a subset of all the modalities used in training.

positioning [1] and radiation dose optimization [2]. In X-ray, the 3D mesh can enable automated radiographic exposure factor selection [3], ensuring optimal radiation dosage to the patient based on patient thickness estimates. Consequently, patient body modeling has seen increasing utility in healthcare [4–7].

Much recent work [8–10] has focused on estimating the 2D or 3D keypoint locations on the patient body. Such keypoints represent only a very sparse sampling of the full 3D mesh in the 3D space that defines the digital human body. The applications noted above necessitate that we go beyond just predicting keypoints and estimate the full 3D mesh representing the patient body. To address this issue, Singh *et al.* [11] presented a technique, using depth sensor data, to retrieve a full 3D patient mesh. However, this method is limited to CT-specific poses and requires depth data. If we change either the application (*e.g.*, X-ray poses and protocols) or even the sensor (*e.g.*, some applications may need RGB-only sensor), this method will need (a) fresh collection and annotation of data, and (b) retraining the model with this new data, both of which may be prohibitively expensive to do repeatedly for each application separately. These issues raise an important practical question: can we design *generic* models that can be trained *just once* and universally used across various scan protocols and application domains? Each application has its own needs and this can manifest in the form of the sensor choice (*e.g.*, RGB-only or RGB-thermal) or specific data scenario (*e.g.*, patient under the cover). To learn a model that can be trained just once and have the capability to be applied across multiple such applications requires what we call *dynamic multi-modal inference* capability. For instance, such a model trained with both RGB and thermal data can now be applied to the following three scenarios without needing any application-specific retraining:

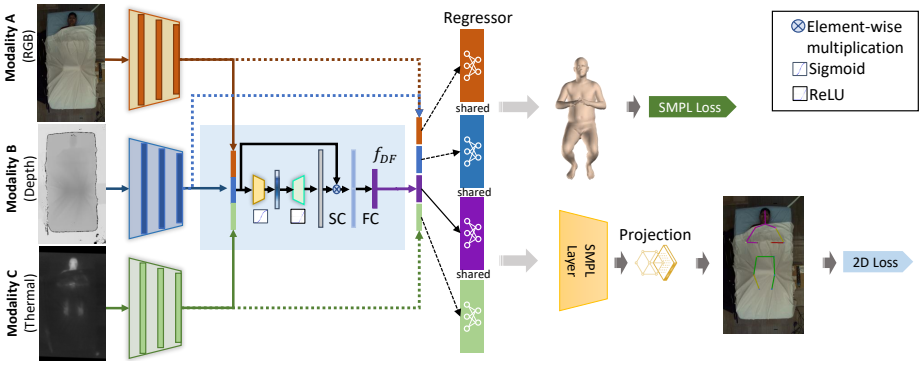

Fig. 2. RDF comprises multiple branches (three shown for illustration) of CNNs to learn a joint multi-modal feature representation, which is used in conjunction with a mesh parameter regressor that outputs the parameters of the 3D patient body mesh.

RGB-only, thermal-only, or RGB-thermal. This ensures flexibility of the trained model to be used in applications that can have an RGB-only sensor, thermal-only sensor or an RGB-thermal sensor. A useful byproduct of such multi-modal inference capability is built-in redundancy to ensure system robustness. For instance, in an application with an RGB-thermal input sensor, even if one of the sensor modalities fails (*e.g.*, thermal stops working), the model above will still be able to perform 3D patient body inference with the remaining RGB-only data. These considerations, however, are not addressed by existing methods, presenting a crucial gap in clinically-deployable and scalable algorithms.

To address the aforementioned issues, we present a new *robust dynamic fusion* (RDF) algorithm for 3D patient body modeling. To achieve the multi-modal inference capability discussed above, RDF comprises a multi-modal data fusion strategy along with an associated training policy. Upon training, our RDF model can be used for 3D patient body inference under any of the possible multi-modal data modality combinations. We demonstrate these aspects under two different two-modality scenarios: RGB-depth and RGB-thermal. In both cases, we evaluate on clinically-relevant patient positioning datasets and demonstrate efficacy by means of extensive experimental comparisons with competing methods.

2 Method

The proposed *robust dynamic fusion* (RDF) framework for 3D patient body modeling comprises several key steps, as summarized in Figure 2. Given multi-modal data input, RDF first generates features in a *joint* multi-modal feature space. While our discussion below assumes two modalities, RDF can be extended to many more modalities as well (Figure 2 shows the scenario with three modalities). Furthermore, to make RDF robust to the absence of any particular modality during testing, we present a probabilistic scheme to perturb the input

data at various multi-modal permutation levels. Our hypothesis with this training policy is that the resulting model will have been trained to predict the 3D patient model even in the absence of any particular input data modality (*e.g.*, if the thermal sensor breaks down, leading to the availability of only the RGB modality data).

Given inputs $\mathbf{I}_{m_1}$ and $\mathbf{I}_{m_2}$ from two modalities m_1 (*e.g.*, RGB) and m_2 (*e.g.*, thermal), RDF first generates feature representations for each modality $\mathbf{f}_{m_1}$ and $\mathbf{f}_{m_2}$ with two separate branches of convolutional neural networks (CNN). These individual feature vectors are then fused with our dynamic feature fusion module to give the feature representation $\mathbf{f}_{\text{DF}}$ of $\mathbf{I}_{m_1}$ and $\mathbf{I}_{m_2}$ in the joint multi-modal feature space. Given $\mathbf{f}_{\text{DF}}$, RDF generates the parameters of the 3D mesh that best describe (as measured by an objective function $L^{\text{DF}}_{\text{mesh}}$ on the mesh parameters) the patient in the input data. These parameters are then used in conjunction with an image projection operation to predict the 2D keypoints, whose error is penalized by means of an objective function L^{DF}_{2D} measuring distance to ground-truth keypoints. To strengthen the representation capability of features in each modality, RDF also computes mesh parameters directly from each of $\mathbf{f}_{m_1}$ and $\mathbf{f}_{m_2}$, each of which are penalized with objective functions $(L^{m_1}_{\text{mesh}}, L^{m_1}_{2D})$ and $(L^{m_2}_{\text{mesh}}, L^{m_2}_{2D})$ respectively. RDF is then trained with the overall loss function:

$$L = L^{\text{DF}}_{\text{mesh}} + L^{\text{DF}}_{2D} + \sum_{i=1}^{M} \left(L^{m_i}_{\text{mesh}} + L^{m_i}_{2D} \right) \tag{1}$$

where M represents the number of input modalities ($M = 2$, *e.g.*, RGB and thermal, in the context above). Note that our proposed approach is substantially different than existing state-of-the-art mesh estimation methods such as HMR [12]. While HMR also regresses mesh parameters from feature representations, it shares the same limitation as Singh *et al.* [11], *i.e.*, it can be trained only for one modality. Consequently, even if one were to use HMR in a multi-modal scenario, it would have to be in a standard two-branch fashion that assumes the availability of data from both modalities during both training and testing, leading to the same limitations and considerations discussed in Section 1. We next discuss each component of our RDF approach in greater detail.

Multi-modal training. To ensure multi-modal inference flexibility discussed above, given $\mathbf{I}_{m_1}$ and $\mathbf{I}_{m_2}$, during training, we simulate several inference-time scenarios with a probabilistic data and training policy, which we achieve by adding noise to our input data streams probabilistically. Specifically, we randomly select one of the two streams m_1/m_2 with a probability p, and replace the input data array of this stream with an array of zeros. With this strategy, as training progresses, the model will have observed all the following three modality possibilities: m_1 only ($\mathbf{I}_{m_2}$ set to zero), m_2 only ($\mathbf{I}_{m_1}$ set to zero), and both m_1 and m_2, thereby "teaching" the model how to infer under any of these scenarios. Given $\mathbf{I}_{m_1}$ and $\mathbf{I}_{m_2}$ (with or without the zero changes as described above), we first extract their individual feature representations $\mathbf{f}_{m_1}$ and $\mathbf{f}_{m_2}$ with their corresponding CNN branches. We then concatenate these two feature vectors,

giving $\mathbf{f}_{\text{cat}}$. Inspired by [13], we process $\mathbf{f}_{\text{cat}}$ with our feature fusion module. This fusion operation, also shown in Figure 2, essentially generates a new feature representation, $\mathbf{f}_{DF}$, that captures interdependencies between different channels and modalities of the input feature representation. Specifically, through a series of fully connected and non-linear activation operations, we produce a vector $\mathbf{sc}$ which can be thought of as a vector of weights highlighting the importance of each channel in the input feature vector $\mathbf{f}_{\text{cat}}$. We then element-wise multiply $\mathbf{f}_{\text{cat}}$ and $\mathbf{sc}$, which is then followed by one more fully connected unit to give $\mathbf{f}_{DF}$.

Mesh recovery. Given $\mathbf{f}_{DF}$, RDF comprises a mesh parameter regressor module (a set of fully connected units) that estimates the parameters of the 3D patient mesh (we use Skinned Multi-Person Linear (SMPL) [14]). SMPL is a statistical model parameterized by shape $\beta \in \mathbb{R}^{10}$ and pose parameters $\theta \in \mathbb{R}^{72}$. The mesh parameter regressor module takes $\mathbf{f}_{DF}$ as input and produces the parameter estimates $\hat{\theta}$ and $\hat{\beta}$, which are penalized by an l_1 distance loss with the ground-truth parameters θ and β:

$$L_{\text{mesh}} = \left\| [\theta, \beta] - [\hat{\theta}, \hat{\beta}] \right\|_1 \tag{2}$$

Keypoints estimation. To ensure accurate estimation of keypoints on the image, our method projects the 3D joints from the estimated mesh to image points. This is achieved using a weak-perspective projection operation [12] that consists of a translation $\rho \in \mathbb{R}^2$ and a scale $t \in \mathbb{R}$. The 2D keypoints are then computed as $\hat{\mathbf{x}}_i = s \prod(\mathbf{X}_i) + \rho$, where $\mathbf{X}_i$ is the i^{th} 3D joint. We then supervise these predictions using an l_1 loss:

$$L_{2D} = \sum_i \left\| \mathbf{x}_i - \hat{\mathbf{x}}_i \right\|_1 \tag{3}$$

where $\mathbf{x}_i$ is the corresponding 2D ground truth.

3 Experiments

Preliminaries. As noted previously, our proposed RDF framework can in-principle be used with any number of input modalities (we only need to increase the number of input streams in Figure 2). However, for simplicity, we demonstrate results with two separate two-modality scenarios: (m_1 = RGB, m_2 = thermal) and (m_1 = RGB, m_2 = depth). In each case, we empirically show the flexibility of RDF in inferring the 3D patient body when any subset of (m_1, m_2) modalities is available at test time. To evaluate the performance of our proposed RDF algorithm, we compare it to a competing state-of-the-art mesh recovery algorithm, HMR [12]. Note that the crux of our evaluation is in demonstrating RDF's flexibility with multi-modal inference. HMR, by design, can be used with only one data modality at a time. Consequently, the only way it can process two data modalities is by means of a two-stream architecture with data from both modalities as input. For this two-stream HMR, note that we use the concatenated features to regress the mesh parameters.

SLP	Train	Test	2D MPJPE ↓	3D MPJPE↓
HMR[12]	RGB	RGB	37.2	155
	T	T	34.2	149
	RGB-T	RGB-T	34.1	143
RDF	RGB-T	RGB	36.6	144
		T	34.7	138
		RGB-T	**32.7**	**137**

SCAN	Train	Test	2D MPJPE↓	3D MPJPE↓
HMR[12]	RGB	RGB	25.6	168
	D	D	23.7	150
	RGB-D	RGB-D	21.8	144
RDF	RGB-D	RGB	17.8	117
		D	21.6	116
		RGB-D	**16.2**	**103**

CAD	Train	Test	2D MPJPE ↓	3D MPJPE↓
HMR[12]	RGB	RGB	7.9	120
	D	D	9.2	118
	RGB-D	RGB-D	6.7	103
RDF	RGB-D	RGB	6.1	106
		D	7.2	104
		RGB-D	**5.7**	**97**

PKU	Train	Test	2D MPJPE↓	3D MPJPE↓
HMR[12]	RGB	RGB	8.8	127
	D	D	13.2	150
	RGB-D	RGB-D	8.2	118
RDF	RGB-D	RGB	7.7	123
		D	11.8	133
		RGB-D	**8.1**	**106**

Table 1. Results on SLP, SCAN, CAD, and PKU. "T": thermal, "D": depth.

SLP	Train	Test	2D MPJPE ↓	3D MPJPE↓
RDF	RGB-D-T	RGB	37.7	144
		T	35.5	135
		RGB-T	**34.0**	**138**

CAD	Train	Test	2D MPJPE ↓	3D MPJPE↓
RDF	RGB-D-T	RGB	6.7	108
		D	7.0	107
		RGB-D	**5.9**	**93**

Table 2. Results on SLP and CAD with three mod. "T": thermal, "D": depth.

Datasets, implementation details, and evaluation metrics. We use the SLP [10] dataset with images of multiple people lying on a bed for the RGB-thermal experiments. These images correspond to 15 poses collected under three different cloth coverage conditions: uncover, "light" cover (referred to as cover1), and "heavy" cover (cover2). We use PKU [15], CAD [16] and an internally-collected set of RGB-D images from a medical scan patient setup (SCAN) for the RGB-depth experiments. PKU and CAD contain a set of complex human activities recorded in daily environment, whereas the SCAN dataset has 700 images of 12 patients lying on a bed in 8 different poses. For SCAN, we create an equal 350-image/6-patient train and test split, and follow the standard protocol for other datasets. In the RDF pipeline, both modality-specific encoder networks are realized with a ResNet50 [17] architecture, which, along with the mesh parameter regressor network, is pretrained with the Humans3.6M dataset [18]. We set an initial learning rate to 0.0001, which is multiplied by 0.9 every 1,000 iterations. We use the Adam optimizer with a batch size of 64 (input image size is 224×224) and implement all code in PyTorch. All loss terms in our objective function have an equal weight of 1.0. For evaluation, we use standard metrics [18]: 2D mean per joint position error (MPJPE) in pixels and 3D MPJPE in millimeters.

Bi-modal inference evaluation. Table 1 shows RGB-T results on the SLP dataset and RGB-D results on the SCAN, CAD, and PKU datasets. In the "HMR" row (in both tables), "RGB" indicates training and testing on RGB-only data (similarly for thermal "T" and depth "D"). The "RGB-T" (and similarly "RGB-D") row indicates a two-stream baseline with the two modality data

Test modality	RGB			Thermal			RGB-T		
Cover condition	uncover	cover1	cover2	uncover	cover1	cover2	uncover	cover1	cover2
HMR[12]	139	150	154	145	149	151	141	145	143
RDF	**137**	**146**	**150**	**135**	**138**	**140**	**134**	**137**	**141**

Table 3. 3D MPJPE (mm) results of SLP evaluation under different cover scenarios.

streams as input. On the other hand, our proposed algorithm processes, during training, both RGB and thermal (or depth) streams of data. However, a key difference between our method and the baseline is how these algorithms are used in inference. During testing, HMR can only process data from the same modality as in training. On the other hand, RDF can infer the mesh with any subset of the input training modalities. One can note from the RGB-T results that RDF with RGB data (144mm 3D MPJPE) is better than the baseline (155mm 3D MPJPE) since it has access to the additional thermal modality data, thereby improving the inference results with the RGB-only modality. A similar observation can be made for the performance comparison on thermal data. RDF (137mm) performs better than the baseline (143mm) in the RGB-T scenario as well, substantiating the role of our feature fusion operation. Similar observations can be made from the evaluation on the SCAN/CAD/PKU datasets too.

Tri-modal inference evaluation. We also evaluate our method with three modalities- RGB, depth, and thermal (RGB-D-T). Since aligned and annotated RGB-D-T data is not available, we instead use our multi-modal training policy to train with pairs of RGB-D and RGB-T data by combining the RGB-T dataset (SLP) with one RGB-D dataset (CAD). The results are shown in Table 2, where one can note our three-branch model is quite competitive when compared to the corresponding separately trained two-branch baselines (3D MPJPE of 93 mm vs. two-branch RDF 97 mm on CAD RGB-D data, 138 mm vs. two-branch RDF 137 mm on SLP RGB-T data).

Under-the-cover evaluation. In Table 3, we evaluate the impact of patient cloth coverage on the final performance. To this end, we use "uncover", "cover1", and "cover2" labels of SLP dataset and report individual performance numbers. One can note that increasing the cloth coverage generally reduces the performance, which is not surprising. Furthermore, since there is only so much information the RGB modality can access in the covered scenarios, as opposed to the thermal modality, the performance with RGB data is also on the lower side. However, RDF generally performs better than the baseline across all these conditions, providing further evidence for the benefits of our method. Finally, some qualitative results from the output of our method are shown in Figure 3.

Noise robustness. We also evaluate the noise robustness of RDF and compare to HMR. In this experiment, with probability p, we replace a particular branch's input with an array of zeros, thus simulating the probabilistic absence of any

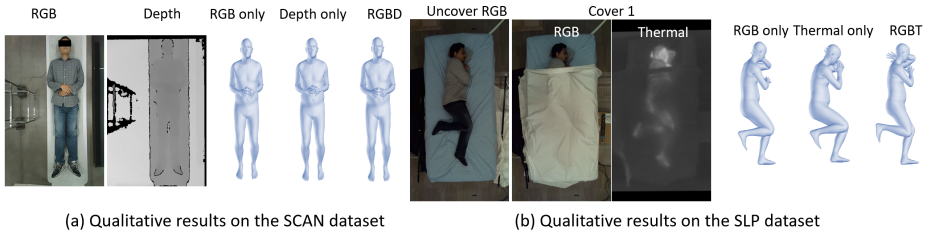

Fig. 3. Qualitative results of the proposed approach on the SCAN and SLP datasets.

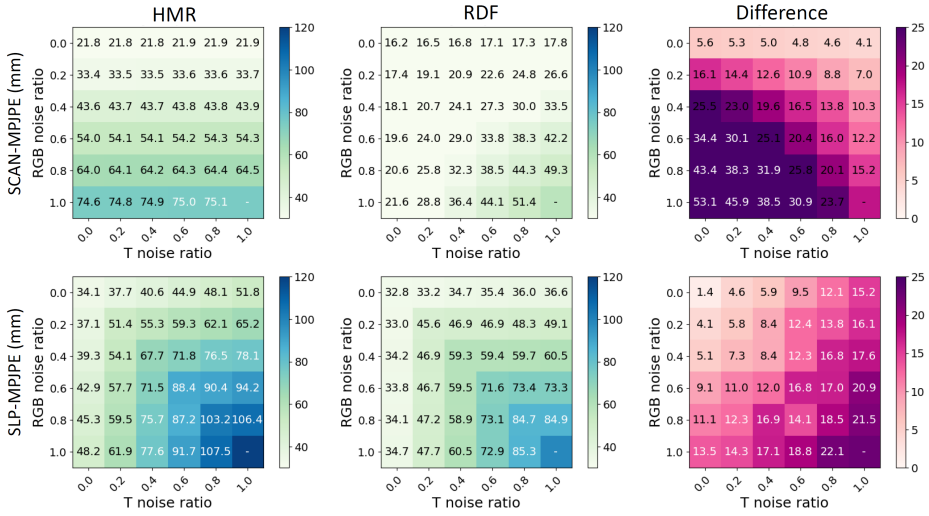

Fig. 4. HMR vs. RDF at various noise levels. "T": thermal, "D": depth.

modality's input during inference (note we ignore the case where both the inputs of both branches are zeros). In Figure 4, we show a matrix representation of the 2D MPJPE of our method as well as baseline HMR, where one can note that with increasing noise level, both methods' performance reduces. Crucially, this performance reduction is lower for our method when compared to the baseline (see difference figure), providing evidence for improved robustness of our method.

4 Summary

We presented a new approach, called robust dynamic fusion (RDF), for 3D patient body modeling. RDF was motivated by a crucial gap of scalability and generality in existing methods, which was addressed by means of RDF's multi-modal inference capability. This was achieved by means of a novel multi-modal fusion strategy, along with an associated training policy, which enabled RDF to infer the 3D patient mesh even when the input at test time is only a subset of

the data modalities used in training. We evaluated these aspects by means of extensive experiments on various patient positioning datasets and demonstrated improved performance compared to existing methods.

References

1. Jianhai Li, Unni K Udayasankar, Thomas L Toth, John Seamans, William C Small, and Mannudeep K Kalra. Automatic patient centering for mdct: effect on radiation dose. *American journal of roentgenology*, 188(2):547–552, 2007.
2. CJ Martin. Optimisation in general radiography. *Biomedical imaging and intervention journal*, 3(2), 2007.
3. William Ching, John Robinson, and Mark McEntee. Patient-based radiographic exposure factor selection: a systematic review. *Journal of medical radiation sciences*, 61(3):176–190, 2014.
4. Leslie Casas, Nassir Navab, and Stefanie Demirci. Patient 3d body pose estimation from pressure imaging. *International journal of computer assisted radiology and surgery*, 14(3):517–524, 2019.
5. Felix Achilles, Alexandru-Eugen Ichim, Huseyin Coskun, Federico Tombari, Soheyl Noachtar, and Nassir Navab. Patient mocap: Human pose estimation under blanket occlusion for hospital monitoring applications. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 491–499. Springer, 2016.
6. Sebastian Bauer, Alexander Seitel, Hannes Hofmann, Tobias Blum, Jakob Wasza, Michael Balda, Hans-Peter Meinzer, Nassir Navab, Joachim Hornegger, and Lena Maier-Hein. Real-time range imaging in health care: a survey. In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, pages 228–254. Springer, 2013.
7. Supriya Sathyanarayana, Ravi Kumar Satzoda, Suchitra Sathyanarayana, and Srikanthan Thambipillai. Vision-based patient monitoring: a comprehensive review of algorithms and technologies. *Journal of Ambient Intelligence and Humanized Computing*, 9(2):225–251, 2018.
8. Vinkle Srivastav, Thibaut Issenhuth, Abdolrahim Kadkhodamohammadi, Michel de Mathelin, Afshin Gangi, and Nicolas Padoy. Mvor: A multi-view rgb-d operating room dataset for 2d and 3d human pose estimation. *arXiv preprint arXiv:1808.08180*, 2018.
9. Vinkle Srivastav, Afshin Gangi, and Nicolas Padoy. Human pose estimation on privacy-preserving low-resolution depth images. In *MICCAI*, 2019.
10. Shuangjun Liu and Sarah Ostadabbas. Seeing under the cover: A physics guided learning approach for in-bed pose estimation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 236–245. Springer, 2019.
11. Vivek Singh, Kai Ma, Birgi Tamersoy, Yao-Jen Chang, Andreas Wimmer, Thomas O'Donnell, and Terrence Chen. DARWIN: Deformable patient avatar representation with deep image network. In *MICCAI*, 2017.
12. Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018.
13. Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

14. Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.
15. Liu Chunhui, Hu Yueyu, Li Yanghao, Song Sijie, and Liu Jiaying. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017.
16. Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. Human activity detection from rgbd images. In *Workshops at the twenty-fifth AAAI conference on artificial intelligence*, 2011.
17. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
18. Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.