DiSRT-In-Bed: Diffusion-Based Sim-to-Real Transfer Framework for In-Bed Human Mesh Recovery

Supplementary Material

In the supplementary material, we provide additional discussions on synthetic datasets for human mesh recovery(Sec. 1), as well as additional details on data augmentation (Sec. 2.1), loss functions (Sec. 2.2), ablation studies (Sec. 3), and visualization examples (Sec. 4) to further demonstrate the effectiveness of the DiSRT-In-Bed framework.

1. Synthetic Datasets for Human Mesh Recovery

Synthetic datasets are widely used in advancing 3D human mesh recovery by providing large-scale, diverse, and accurately labeled data that would be difficult and expensive to obtain through real-world capture. Prior works such as AGORA [5], BEDLAM [1], and SynBody [7] demonstrate that incorporating synthetic data in training enhances human mesh recovery performance. However, general synthetic datasets are not directly applicable to in-bed scenarios, as lying poses are underrepresented. For in-bed human mesh recovery, BodyPressure [3] builds upon Bodies at Rest [2] to enhance synthetic dataset generation. It leverages physics-based simulation to produce realistic depth and pressure images, better capturing human-bed interactions and occlusions. Additionally, BodyPressure and BodyMAP further demonstrate that scenario-specific synthetic datasets can improve in-bed human mesh estimation.

2. Additional Details about Training Strategy

2.1. Data Augmentation

To enhance the robustness of the diffusion model during training and fine-tuning, we apply various data augmentation techniques to the input depth images for both synthetic and real datasets, simulating complex real-world scenarios. As shown in Fig. 1, the following augmentations are applied:

- **Random Rotation:** Depth images are randomly rotated to introduce variability in human in-bed poses.
- **Random Erase:** Portions of the depth image are randomly masked, simulating occlusions caused by objects such as tables or blankets covering parts of the human body.
- **Random Noise:** Gaussian noise is added to mimic the noise introduced by depth sensors and environmental factors.

These augmentations aim to improve the model's ability to generalize to diverse and challenging real-world conditions.

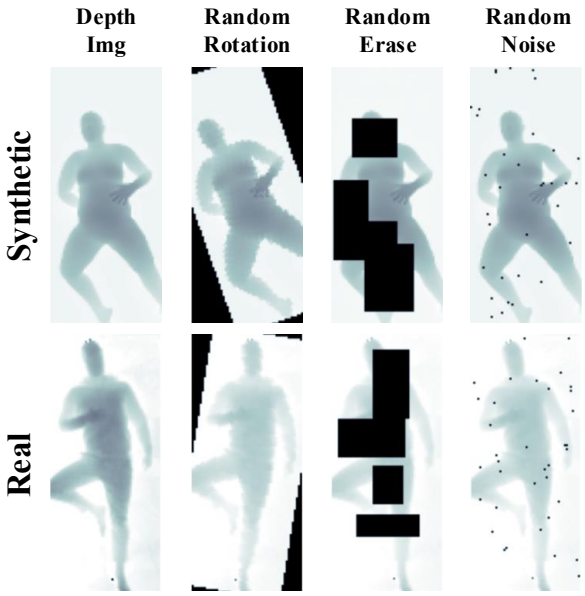


Figure 1. **Illustration of Data Augmentation Operations.**

2.2. Loss Functions

The total diffusion loss used to train and fine-tune the diffusion model contains two components: SMPL parameter loss and vertex position loss. For SMPL parameter loss, we employ standard human pose and shape regularization loss utilized in BodyMAP [6] as follows:

$$\begin{aligned}\mathcal{L}_{\text{SMPL}} &= \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_1 + \lambda_{\boldsymbol{\theta}}\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|_1 \\ &\quad + \lambda_{\psi_{\boldsymbol{x}}}\left(\|\mathbf{u}-\hat{\mathbf{u}}\|_1 + \|\mathbf{v}-\hat{\mathbf{v}}\|_1\right) + \lambda_{\boldsymbol{J}}\sum_{i=1}^{24}\|\mathbf{j}_i-\hat{\mathbf{j}}_i\|_2, \\ \lambda_{\boldsymbol{\beta}} &= \frac{1}{10\sigma_{\boldsymbol{\beta}}}, \quad \lambda_{\boldsymbol{\theta}} = \frac{1}{69\sigma_{\boldsymbol{\theta}}}, \quad \lambda_{\psi_{\boldsymbol{x}}} = \frac{1}{6\sigma_{\psi_x}}, \quad \lambda_{\boldsymbol{J}} = \frac{1}{24\sigma_J},\end{aligned} \tag{1}$$

where each hyper-parameter term is normalized by standard deviations of body parameters $\sigma_{\boldsymbol{\beta}}$, joint angles $\sigma_{\boldsymbol{\theta}}$, continuous global rotation $\sigma_{\psi_{\boldsymbol{x}}}$ and Cartesian joint positions, computed from the entire synthetic training set. $\mathbf{j}_i \subset \mathbf{J}$ represents the Cartesian position of a single joint. Additionally, we use vertex loss to further enhance diffusion stability and performance:

$$\mathcal{L}_{\text{v2v}} = \frac{1}{N_{\mathbf{V}}\sigma_{\mathbf{V}}}\sum_{i=1}^{N_{\mathbf{V}}}\|\mathbf{v}_i-\hat{\mathbf{v}}_i\|_2 \tag{2}$$

where $\mathbf{v}_i \subset \mathbf{V}$ represents the Cartesian position of a single

human mesh vertex, $N_{\mathbf{V}} = 6890$ vertices, and the loss term is normalized by $\sigma_{\mathbf{V}}$.

Thus, the total loss for the diffusion reverse process is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SMPL}} + \lambda_{\text{v2v}}\mathcal{L}_{\text{v2v}}, \tag{3}$$

where λ_{v2v} is a tunable hyper-parameters. We set $\lambda_{\text{v2v}} = 1.0$ for all the experiments.

2.3. Learning Rate Scheduler

As mentioned in Sec. 4.3, we adopt a linearly adjusted learning rate scheduler to adapt to varying amount of real-world data during the fine-tuning stage. Specifically, given the initial learning rate lr_init, the current step index step_cur, and the total number of fine-tuning steps steps_total, the current learning rate is computed as:

$$\text{lr\_cur} = \left(1 - \frac{\text{step\_cur}}{\text{steps\_total} + 1}\right) \text{lr\_init}. \tag{4}$$

3. Additional Ablation Study

3.1. Effectiveness of Loss function

Loss	MJPJE	PVE
SMPL Loss	53.48	66.86
SMPL Loss + v2v Loss	50.81	61.18

Table 1. **Ablation on Loss Function.**

We conduct an ablation study by comparing models trained with different loss functions using the complete synthetic and real training datasets. Tab. 1 shows that adding the v2v loss term to the total loss function enhances the model's performance in mesh estimation in terms of both MPJPE and PVE metrics.

3.2. Additional Comparisons of PVE Results

In addition to the results presented in Sec.5.5 of the main paper, we provide charts for the PVE metric to further demonstrate the effectiveness of our Sim-to-Real Transfer Framework and the proposed diffusion model architecture. The trends observed in PVE results across varying real data utilization percentages align closely with those of the MPJPE results.

Fig. 2a shows that leveraging synthetic data substantially enhances model performance in the PVE metric. Fig. 2b demonstrates that integrating our Sim-to-Real Transfer Framework into the BodyMAP model results in significant improvements, particularly under scenarios with limited access to real-world data. Additionally, Fig. 2c compares four diffusion model designs on the PVE metric. Although the differences in PVE are less pronounced compared to the MPJPE results shown in Fig.5c of the main paper, our proposed architecture consistently outperforms other design choices.

3.3. Effectiveness of Fine-tuning Strategies

In the fine-tuning stage, we introduce a linearly and automatically adjusted scheduler as described in Sec.4.3 of the main paper. The initial learning rate is set to match that of the training stage, i.e., $lr = 1 \times 10^{-4}$. During fine-tuning, the learning rate and weight decay are updated at each diffusion step using the AdamW optimizer. Specifically, for each step, we input a batch of depth images paired with randomly generated timesteps and generate noisy SMPL parameters by iteratively adding Gaussian noise to the ground truth SMPL parameters based on the given timestep. The diffusion model then learns to denoise these SMPL parameters and directly predict the ground truth parameters, as detailed in Sec.4.2.1 of the main paper.

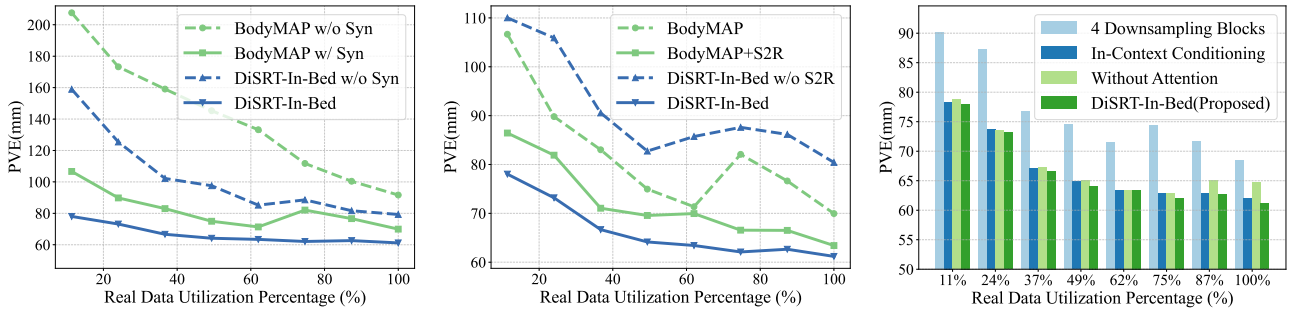
In Fig. 3, we compare the performance of models in terms of MPJPE and PVE across different data splits, using various fine-tuning scheduler strategies, including linear, cosine, exponential, and no scheduler. For the linear and cosine schedulers, the maximum number of iterations depends on the amount of real data available and the number of epochs used for fine-tuning. For the exponential scheduler, we set the multiplicative decay factor for the learning rate to 0.999. The results show that the linearly-adjusted scheduler achieves consistently lower errors compared to other approaches. This demonstrates the effectiveness of our fine-tuning strategy in improving the model's performance.

3.4. Effectiveness of Synthetic Data Utilization

In Table 1, we present experiments using all synthetic data combined with varying proportions of real training data to validate the generalizability and effectiveness of the proposed DiSRT-In-Bed pipeline. Additionally, we perform experiments to further demonstrate the impact of incorporating synthetic data. In this setting, training is conducted using all real data combined with different proportions of synthetic data, while testing is performed on the same real dataset. As shown in Fig. 4, both MPJPE and PVE generally decrease as the proportion of synthetic data increases. However, a slight increase in error metrics is observed when synthetic data reaches 70% and 90% due to distribution shifts. Overall, the best performance is achieved when using all synthetic data and all real training data, as presented in Sec. 5, compared to settings with less synthetic data.

4. Additional Visualizations

We present additional visualization examples to illustrate the effectiveness of our DiSRT-In-Bed method compared to the state-of-the-art BodyMAP method. As shown in Fig. 6, our proposed method achieves superior mesh predictions, especially when access to real-world data is limited. The predictions from our model align more closely with the in-



(a) Ablation on Synthetic Data (b) Ablation on Training Strategies (c) Ablation on Different Architectures

Figure 2. **Additional Ablation Study on Diffusion-Based Sim-to-Real Transfer Framework.**

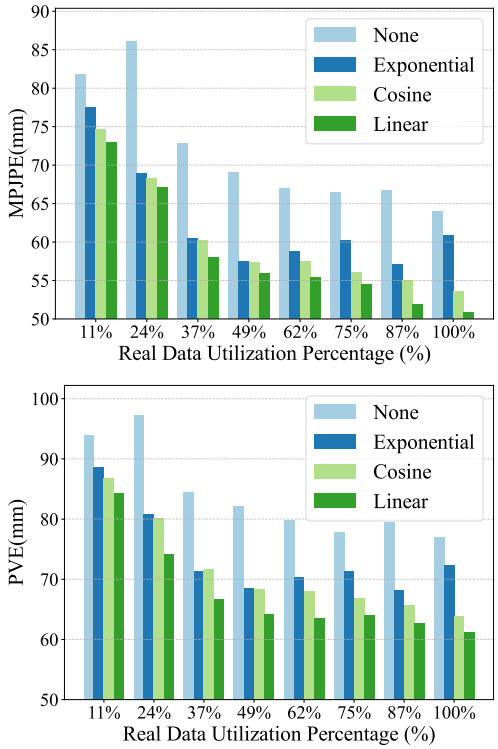


Figure 3. **Ablation Study on Fine-tuning Schedulers.**

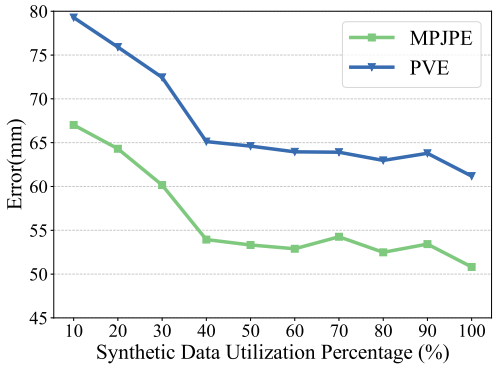


Figure 4. **Ablation Study on Synthetic Data Utilization.**

put data and exhibit stable performance across varying covering scenarios.

Fig. 7 provides additional visualizations on the SLP [4] hospital setting dataset, **which features a different data distribution from the training dataset and lacks labeled ground truth**. Here, we compare our method, with and without the proposed Sim-to-Real training strategies described in Sec.4.3 of the main paper, against BodyMAP in terms of generalization to diverse real-world settings. All models were trained on the complete synthetic dataset and the full real-world SLP [4] home setting dataset.

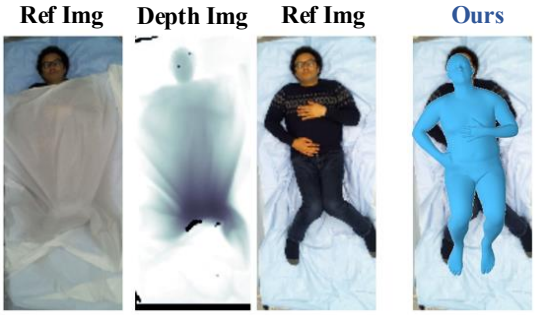
The results reveal that our method without the Sim-to-Real training strategies performs comparably to BodyMAP; however, both are less stable across different covering scenarios and fail to capture finer details. In contrast, our proposed Sim-to-Real framework significantly enhances stability and detail alignment, demonstrating its robustness and generalization capability across varying real-world conditions.

5. Limitations and Future Work

While our proposed DiSRT-In-Bed demonstrates promising performance in handling in-bed human mesh recovery with limited real-world data and strong generalization across different environmental settings, there are two key directions for future work: improving accuracy and enhancing scalability.

Accuracy: Future efforts could focus on improving the prediction quality of in-bed human body meshes. For instance, as shown in Fig. 5a, failure cases involving self-interpenetration remain challenging. In the first example, interpenetration occurs near the left foot and right knee due to the complex pose and the close proximity of these body parts. Similarly, in the second example, self-contact introduces ambiguity in determining the precise position of body parts. Addressing these issues could involve refining model components to better account for self-contact scenarios or

(a) Self-Interpenetration.



(b) Misalignment.

Figure 5. **Failure Cases of DiSRT-In-Bed.**

incorporating additional constraints to reduce interpenetration errors.

Scalability: Extending DiSRT-In-Bed to establish its clinical effectiveness is another critical direction. Fig. 5b highlights a misaligned prediction caused by a challenging, out-of-distribution input from the SLP [4] hospital-setting dataset. Addressing such misalignments in different settings could involve several approaches: expanding synthetic datasets using customizable simulations, incrementally fine-tuning the diffusion model with newly collected data, and designing new diffusion model components that integrate domain-specific knowledge. These advancements could push our framework closer to practical deployment in clinical environments.

References

[1] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, 2023. 1

[2] Henry M Clever, Zackory Erickson, Ariel Kapusta, Greg Turk, Karen Liu, and Charles C Kemp. Bodies at rest: 3d human pose and shape estimation from a pressure image using synthetic data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6215–6224, 2020. 1

[3] Henry M Clever, Patrick L Grady, Greg Turk, and Charles C Kemp. Bodypressure-inferring body pose and contact pressure from a depth image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):137–153, 2022. 1

[4] Shuangjun Liu, Xiaofei Huang, Nihang Fu, Cheng Li, Zhongnan Su, and Sarah Ostadabbas. Simultaneously-collected multimodal lying pose dataset: Enabling in-bed human pose monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1106–1118, 2023. 3, 4, 5, 6

[5] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[6] Abhishek Tandon, Anujraaj Goyal, Henry M Clever, and Zackory Erickson. Bodymap-jointly predicting body mesh and 3d applied pressure map for people in bed. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2480–2489, 2024. 1

[7] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20282–20292, 2023. 1

Figure 6. **Additional Visualization Comparison with Baseline on the SLP [4] Home-Setting Dataset.**

Figure 7. **Additional Visualization Comparison with Baseline on the SLP [4] Hospital-Setting Dataset.**