

Appendix

In this appendix, we provide additional detailed implementations, qualitative comparisons and ablation studies in Section A, Section B and Section C.

A. Detailed Implementations

We utilize Stable-Diffusion [18] V1-5 with ControlNet pre-trained specifically on skeleton image conditions so that we can further unleash its diffusion priors. The denoising U-Net has 25 blocks in total, with a middle block and 12 input and output blocks each. The ControlNet takes 12 input blocks as a parallel branch, connected with zero convolution layers as the output layer. The latent code z_0 is in 32×32 and is downsampled to 16×16 and 8×8 resolutions. During the upsampling process in output blocks, we extract cross-attention maps in the shape of 8×8 and 16×16 and concatenate them with corresponding feature maps in channel dimension. Finally, we use linear layers to merge the pyramid multi-layer feature maps into $2048 \times 8 \times 8$ for subsequent regression processing. For the teacher model, we employ the same framework as the student.

As for the ablation study of various backbones, we follow [2, 13] to initially extract an early-stage image feature in 64×64 . Note that for convolution-based backbones (e.g., ResNet50 [7] and HRNet-W48 [19]), we use a convolution kernel with 2 strides and a max-pooling layer to downsample the image, while for transformer-based backbones (e.g., ViT-L [3] and Swin-V2-L [14]) we apply patch embedding layers. We set the patch size to 4 for Swin-V2-L and 16 for ViT-L, considering the latter’s substantial computational cost. We further quantify the amount of model parameters as illustrated in Table A. With the well-designed implementation of LoRA [8], we significantly reduce the trainable parameters while effectively maintaining the diffusion prior in the pre-trained models. In comparison with ViT-L, our backbone achieves superior results with reduced computational expenses.

We apply a pose parameter decoder pre-trained on a huge SMPL dataset AMASS [15]. The framework of our VQ-VAE is built with linear layers and we take the pose parameters in rotation matrix representation as input in the shape of 24×9 . The codebook class number is 2048 and the token dim is 256. During the pre-train stage, we supervise the results with the reconstruction loss following [17]. We also utilize Exponential Weighted Average on codebook optimization inspired by [6]. As for the inference stage, the regressor will provide 48 tokens to the decoder and finally retrieve the SMPL pose parameters $\Theta \in \mathbb{R}^{24 \times 3}$.

For training loss, we set λ_{2D} , λ_{3D} , λ_{SMPL} to 5.0, 2.0 and 1.0 respectively. λ_{NKR} is set to 0.1. We speed up training by using distributed training with Pytorch [16] using 8 Nvidia GeForce RTX 4090 GPUs.

Table A. **Comparison on model parameters.** We quantify the amount of total parameters and trainable parameters across different backbone configurations in the entire pipeline. Given the implementation of LoRA [8], our DPMesh achieves remarkable results with a minimal fraction of weights finetuned, demonstrating its efficiency and effectiveness.

Backbone	Total Params.	Trainable Params.	MPIPE↓
ResNet50 [7]	39.7M	39.0M	76.1
HRNet-W48 [19]	77.8M	77.1M	75.6
ViT-L [3]	1257.2M	1256.4M	73.1
Swin-V2-L [14]	211.2M	210.3M	73.5
DPMesh (Ours)	1426.3M	408.9M	70.9

B. More Qualitative Results

Robustness to noisy 2D key-points. We illustrate the prediction of DPMesh under noisy key-points compared with previous methods in Figure A. We draw the bounding box according to the region encompassing visible key-points. Given the complexities associated with individual interactions and the absence of precise key-point hints, traditional approaches may yield false predictions. However, by incorporating a robust diffusion backbone and the Noisy Key-point Reasoning (NKR) approach, our DPMesh algorithm achieves a marked improvement in accuracy.

Comparison on OCHuman [22]. We present additional qualitative assessments on OCHuman, an in-the-wild dataset with substantial occlusion consisting of 8,110 meticulously annotated human instances across 4,731 images. Initially, we employ AlphaPose [4] as an off-the-shelf detector to obtain coarse 2D key-points. Subsequently, we estimate the mesh results with previous methods and our DPMesh. As shown in Figure B, our DPMesh demonstrates exceptional performance in tackling challenging occlusions and complex human poses.

Comparison on CrowdPose [10]. The CrowdPose dataset comprises 8,000 images characterized by dense occlusions and complex crowd scenarios. We compare our DPMesh with 3DCrowdNet [2], which is tailored for in-the-wild crowded scenes and addresses 3D human mesh recovery issues with a joint-based regressor. As shown in Figure C, our DPMesh proficiently estimates the shape and pose of all individuals within the view, effectively handling ambiguous person interactions and body truncations.

C. Additional Ablation Studies

Type of 2D spatial conditions. In conventional controllable generative models such as ControlNet [21], the input human pose guidance is typically provided in the form of a RGB skeleton image detected from Openpose [1]. Adhering to the approach of ControlNet, we first extract image features from the skeleton image using convolutional layers



Figure A. **Illustration of mesh recovery results under noisy 2D key-points.** In occlusion and crowded scenes, there are 2D key-points missed or inaccurately predicted. Compared with previous methods, our DPMesh with a well-designed Noisy Key-point Reasoning (NKR) approach successfully recovers the correct human mesh.

and then concatenate the spatial feature with z_0 as a spatial condition. As summarized from Table B, we assume that the heatmap guidance carries more information such as the joint correspondence to different heatmap channels, than a single skeleton image. Furthermore, noisy key-points may lead to incorrect skeleton connections, which can neg-

atively impact performance. Therefore, our DPMesh utilizes heatmap guidance to effectively introduce spatial conditions.

Implementation of different mesh regressors. In order to assess the efficacy of diffusion-based backbone, we apply recurrent linear layers (RLL) derived from [9] and cas-



Figure B. **Qualitative comparison on OCHuman dataset [22].** In more challenging occlusion and crowded scenarios, our DPMesh precisely predicts the human mesh, effectively disregarding interference from adjacent individuals.

Table B. **Ablation study of spatial condition type.** We study the impact of different spatial guidance to denoising U-Net.

Conditions	3DPW		3DPW-OC	
	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓
skeleton map	74.9	48.4	72.1	49.7
heatmap	73.6	47.4	70.9	48.6

cade transformer decoder learned from [13] as the SMPL regressor. Results are presented in Table C. Without bells and whistles, even with vanilla recurrent linear layers, our diffusion-based feature extractor outperforms the performance of JOTR [13], which carefully designs a con-

Table C. **Ablation study of different SMPL regressors.** We deploy two distinct types of SMPL regressors (*e.g.*, recurrent linear layers (RLL) [9] and cascade transformer decoder (CTD) [13]) on both ResNet50 [7] and our diffusion-based backbone to investigate their respective performance contributions.

Settings	3DPW		3DPW-OC	
	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓
ResNet+RLL	81.6	53.9	82.2	54.1
ResNet+CTD	80.2	52.4	78.9	52.8
Diffusion+RLL	75.4	49.0	72.5	50.0
Diffusion+CTD	73.6	47.4	70.9	48.0



Figure C. **Qualitative comparison on CrowdPose dataset [10].** We compare DPMesh with 3DCrowdNet which excels in managing crowded environments. Our DPMesh yields superior outcomes in complex situations such as person-person ambiguity, occlusion with camera distortion and intricate poses within crowds.

Table D. **Ablation study of LoRA [8] implementation.** We adjust the LoRA rank hyperparameter and unlock more frozen layers in pre-trained diffusion model (*e.g.*, ✓: Weights are unlocked with learnable LoRA matrices. ✗: Weights remain consistent with the pre-trained model.). Results are evaluated on 3DPW-OC [20, 23].

LoRA rank	ResBlock	MPJPE↓	PA-MPJPE↓
8	✗	73.2	50.1
64	✓	75.9	51.5
64	✗	70.9	48.0

trastive learning loss for its cascade transformer decoder. These findings indicate that our outstanding performance is not severely dependent on the specific regressor employed, highlighting the versatility and effectiveness of the diffusion-based backbone.

Influence of LoRA [8]. In order to preserve the diffusion

Table E. **Ablation study on occluded parts.** We conduct an ablation study on our new metrics with cross-attention module (CAM) and Noisy Key-point Reasoning (NKR).

Settings	MPJPE↓	PA-MPJPE↓	OCC-MPJPE↓	OCC-PA-MPJPE↓
JOTR [27]	75.7	52.2	89.2	63.5
w/o CAM	73.7	50.6	89.4	61.0
w/o NKR	73.9	49.9	91.1	61.6
DPMesh	70.9	48.0	86.2	57.6

prior within the pre-trained model and minimize computational expenses, we utilize LoRA to finetune only a few parameters in U-Net. We study different LoRA ranks and unlock more blocks to optimize. As shown in Table D, lower LoRA rank fails to thoroughly unleash the potential of diffusion for visual perception tasks and further unfreezing the ResBlocks may compromise the diffusion prior learned from extensive data. Therefore, employing LoRA matrices

Table F. **Comparison with more methods.** We finetune DPMesh with 3DPW training set and compare it with more SOTA methods. DPMesh achieves a competitive result on 3DPW and significantly outperforms them on occlusion benchmark.

Method	3DPW		3DPW-OC	
	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓
HyBrIK [11]	71.6	41.8	90.8	58.8
NIKI [12]	71.3	40.6	85.5	53.5
DPMesh-ft	68.4	42.8	70.9	48.0

Table G. **Comparison with HMDiff.** Compared with step-by-step diffusion framework, DPMesh exhibits superior performance, particularly on the 3DPW-PC benchmark.

Method	3DPW		3DPW-PC	
	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓
HMDiff [5]	72.7	44.5	114.2	73.5
DPMesh (ours)	68.4	42.8	82.2	56.6

to simply unlock cross-attention blocks is appropriate for this specific task, striking a balance between performance and computational efficiency.

Evaluation on occluded joints. Our NKR focuses on dealing with noisy guidance from erroneous key-points. Given that these noisy key points only take a small fraction of the total body, they may not have a significant impact on the overall result. To further assess the NKR’s impact, we introduce **OCC-MPJPE↓** and **OCC-PA-MPJPE↓** to evaluate errors on occluded joints. As shown in Table E, the cross-attention module and the Noisy Key-point module both are effective on occluded key-points input.

More comparisons. We compare DPMesh with more methods. For a fair comparison, when testing on the 3DPW test split, we fine-tune our model with the 3DPW [20] training set. As illustrated in Table F, DPMesh achieves competitive performance on the 3DPW test split benchmark and significantly outperforms the occlusion benchmark. Furthermore, we compare our one-step DPMesh with the step-by-step denoising framework HMDiff [5]. HMDiff is an optimization method that takes over 200 steps to recover human mesh. As shown in Table G, DPMesh exhibits much better results on the 3DPW-PC benchmark while also outperforming HMDiff on the 3DPW test split.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, 2017. 1

[2] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *CVPR*, pages 1475–1484, 2022. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov,

Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 1

[4] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alpha-pose: Whole-body regional multi-person pose estimation and tracking in real-time. *TPAMI*, 2022. 1

[5] Lin Geng Foo, Jia Gong, Hossein Rahmani, and Jun Liu. Distribution-aligned diffusion for human mesh recovery. In *ICCV*, pages 9221–9232, 2023. 5

[6] Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. Human pose as compositional tokens. In *CVPR*, 2023. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 3

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 1, 4

[9] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *CVPR*, pages 2252–2261, 2019. 2, 3

[10] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, pages 10863–10872, 2019. 1, 4

[11] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, pages 3383–3393, 2021. 5

[12] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *CVPR*, pages 12933–12942, 2023. 5

[13] Jiahao Li, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Jotr: 3d joint contrastive learning with transformers for occluded human mesh recovery. In *ICCV*, pages 9110–9121, 2023. 1, 3

[14] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, pages 12009–12019, 2022. 1

[15] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 1

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. 1

[17] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 1

- [18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. [1](#)
- [19] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. [1](#)
- [20] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. [4](#), [5](#)
- [21] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. [1](#)
- [22] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *CVPR*, pages 889–898, 2019. [1](#), [3](#)
- [23] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020. [4](#)