- **Shared Questions**

Are methods reported in Table 1 (main paper) trained with the same DeepHuman dataset?

Yes, using their GitHub code. We will also release training /test/evaluation scripts of these competing methods. Moreover, we include results of pre-trained models released by PIFu and PIFuHD in Table 1. Under the same training data, PIFuHD achieves lower relative improvement over PIFu than Geo-PIFu. More discussions on PIFuHD are below.

Table 1: DeepHuman benchmarks. Parameter size of Geo-PIFu is 30616954 (*12 times smaller than PIFuHD*).

Method	Parameter Size	Mesh		Normal	
		CD	PSD	Cosine	L2
PIFu	15604738	10.571	9.285	0.1422	0.4141
PIFuHD	387049625	9.489	9.349	0.1228	0.3776

Why not conduct detailed comparisons with PIFuHD in the main paper?

1) PIFuHD was published at CVPR 2020 after the NeurIPS 2020 submission deadline.
2) Such a comparison would be unfair. Although not emphasized in their paper, PIFuHD uses ImageNet for pre-training, additional networks, and higher resolution inputs than all other competing methods and than our method.
3) In the limited time available to rebut, we show pre-trained PIFuHD results in Table 1 by upsampling 512x512 images to 1024x1024 (their code/model). PIFuHD has not released the training code, which involves several stages for multiple networks. For example: one stack-hourglass for global-PIFu, one stack-hourglass for fine-PIFu, one customized ResNet for front normal, one customized ResNet for back normal, and one ImageNet-pretrained VGG-16 for perceptual losses. PIFuHD uses much heavier networks than Geo-PIFu, and requires complex training steps (end-to-end training is even worse than PIFu). In contrast, we provide all training/test/evaluation scripts to make Geo-PIFu fully reproducible. □
4) The two ideas of PIFuHD (using sliding windows to ingest high resolution images, and offline estimated front/back normal maps to further augment input color images) are both add-on modules *wrt.* (Geo)-PIFu. Given high resolution images and offline estimated normal maps, one might combine PIFuHD with Geo-PIFu for further improved local surface details and global topology robustness. But this is out of the scope of our work.

- **Reviewer #1**

Computation considerations. Please see Table 1 and answers □, ◯ for more discussions on computation cost of PIFuHD and Geo-PIFu. Real-time performance is a common challenge for concurrent works, e.g. PIFuHD, ARCH.

The advantage of 3D Decoder is not quantitatively described. These results are in *exp-a* of Table 2 (main paper).

- **Reviewer #2**

More details on the late fusion experiment in Table 2 (main paper). We add 3 FC layers (112, 224, 256) with Leaky-ReLU after obtaining the geometry-aligned features for late fusion. These layers introduce additional computation cost than the early fusion method, which we use in our benchmarks and qualitative demos for its good balance of computation and performance. More discussions on the capacity of latent voxel features are in answers △ and ◯.

Quantitative results of the 3D GAN loss. Mesh: CD (2.464), PSD (3.372). Normal: Cosine (0.1298), $L2$ (0.4148). We did not include them because they are not the main focus of our work. We will add these results in camera ready.

Parametric body models. (Accuracy) Current methods suffer from large pose errors, hurting the rest reconstruction steps. Thus, DeepHuman has large CD/PSD. (Computation) Although not emphasized in ARCH/DeepHuman papers, parametric shape estimation networks that they rely on involve many times more computation cost than the rest modules.

More discussions on IF-Net, CVPR 2020. While IF-Net takes partial or noisy 3D voxels as input, Geo-PIFu only utilizes a single-view color image. Thus, IF-Net has access to "free" 3D shape cues of the human subject. But Geo-PIFu must achieve an ill-posed 2D to 3D learning problem. Meanwhile, Geo-PIFu needs to factorize out pixel domain nuisances (*e.g.* colors, lighting) in order to robustly recover the underlying dense/continuous occupancy fields. △

- **Reviewer #3**

More results on other datasets like BUFF and DeepFashion. The *BUFF test data comprises only 5 front-facing images with simple poses and no self-occlusion*; comparison would not add significant new insight to the existing evaluation comprising 21744 test images of various poses, camera angles and lighting. We will add visual results on DeepFashion in camera ready.

- **Reviewer #4**

Data preparation. a) Following DeepHuman, we use OpenDR with Lambertian point lights for image rendering. We will release the rendering scripts. We saved camera pose and lighting settings of each image so that our data can be reproduced. b) and c) We use the same point sampling strategy and data normalization method as PIFu. Therefore we can fairly evaluate the impact of our proposed modules. Query points sampling is a critical process that deserves in-depth studies as a full paper, *e.g.* one emerging work is Curriculum DeepSDF (Duan, Yueqi, et al. ECCV 2020).

Study of global feature. The latent voxel feature resolution: (C-8, D-32, H-48, W-32), in total 393216. In comparison, the latent pixel feature resolution: (C-256, H-128, W-128), in total 4194304. Studying different resolutions of the latent voxel features is very interesting. We promise to add this experiment in camera ready to make our paper stronger. ◯

Limited discussions and incremental contribution. We will explore the directions mentioned in additional feedback in our future work. As pointed out by the reviewer, many problems are common, open challenges for concurrent works.