Point Voxel Bi-directional Fusion Implicit Field for 3D Reconstruction: Supplemental Materials

Chuanmao Fan*
cf7b6@missouri.edu
University of Missouri
Columbia, Missouri, USA

Chenxi Zhao†
chenxiz@clemson.edu
Clemson University
Clemson, South Carolina, USA

Kevin Xue
kzx2n8@missouri.edu
University of Missouri
Columbia, Missouri, USA

Ye Duan‡
duan@clemson.edu
Clemson University
Clemson, South Carolina, USA

ACM Reference Format:
Chuanmao Fan, Chenxi Zhao, Kevin Xue, and Ye Duan. 2024. Point Voxel Bi-directional Fusion Implicit Field for 3D Reconstruction: Supplemental Materials. In *GI'24: Graphics Interface 2024, June 03–06, 2024, Halifax, Nova Scotia, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 NETWORK AND TRAINING

Platform and hardware: The proposed Bifusion framework is implemented with Pytorch [7]. The training and testing are conducted using a middle-range desktop computer with an Nvidia RTX 4090 GPU of 24 GB memory.

Network architecture:The volume branch of Bifusion is constructed with a five layers U-shaped network. And similarly, the point branch is built with five layers of point based U-net. Volume and points modules' design are mirror symmetric to facilitate point and voxel feature fusion. Their feature dimensions are [4, 16, 32, 64, 128, 128, 64, 32, 16, 16] from encoder to decoder of the U-shaped net.

Fusion design: We implemented two versions of bidirectional fusion modules. The self-contained figure 1 demonstrate the design for side by side comparison. Figure 1(a) illustrate the design used in ablation study Bifusion v1. The design shown in Figure 1(b) corresponds to Bifusion v2 which is mainly tested in our work.

Loss functions: We use binary cross entropy loss of Equation 1 for occupancy learning:

$$L_o\left(W\right) = \frac{1}{|B|}\sum_{i=1}^{|B|}\sum_{j=1}^{K}\left|BCE\left(O_{q,i,j}, O_i^j\right)\right| \tag{1}$$

Here B is the mini-batch data size, K is the number of query points for each object, $O_{q,i,j}$ is the prediction value for a given query point q_i^j, BCE is the binary cross entropy loss for occupancy field.

Training: The network is trained using the Adam optimizer[4] with parameters $\beta_1 = 0.9, \beta_2 = 0.999$, and an initial learning rate of $1.0e^{-4}$. The learning rate decreases by $0.1\times$ with the step scheduler at 50 and 100 epochs, respectively. We use the same ratio of 7:2:1 for training, validation, and testing for all the datasets.

Metrics: **Chamfer distance (CD)** as the metric for performance evaluation More specifically, we sample points on both the reconstruction and the ground truth surface to serve as the proxy for computing the chamfer distance between the two surfaces. The chamfer distance between the two shapes represented by point cloud P_a and P_b respectively can thus be measured as the sum of the average of the minimum distances from P_a to P_b and from P_b to P_a. In the paper, we follow ONet [6], IFnet [2], we compute both $CDl1$ and $CDl2$.

$$\begin{aligned} Chamferl1(P_a, P_b) &= \frac{Completeness}{2|P_a|} + \frac{Accuracy}{2|P_b|} \\ Chamferl2(P_a, P_b) &= \frac{Completeness^2}{2|P_a|} + \frac{Accuracy^2}{2|P_b|} \end{aligned} \tag{2}$$

where

$$\begin{aligned} Completness &= \sum_{p_a \in P_a} \min_{p_b \in P_b} d(p_a, p_b) \\ Accuracy &= \sum_{p_b \in P_b} \min_{p_a \in P_a} d(p_b, p_a) \\ Completness^2 &= \sum_{p_a \in P_a} \min_{p_b \in P_b} d(p_a, p_b)^2 \\ Accuracy^2 &= \sum_{p_b \in P_b} \min_{p_a \in P_a} d(p_b, p_a)^2 \end{aligned} \tag{3}$$

Normal Consistency (NC). The normal consistency between two points cloud P_a and P_b is defined by the following equation:

*Co-first author.
†Co-first author.
‡Corresponding author.

Graphics Interface 2024, June 03–06, 2024, Halifax, Nova Scotia

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

$$NC(P_a, P_b) = \frac{1}{2|P_a|} \sum_{p_a \in P_a} N_{p_a} N_{near p_a, P_b} + \frac{1}{2|P_b|} \sum_{p_b \in P_b} N_{p_b} N_{near p_b, P_a} \tag{4}$$

where $N_{near p_a, P_b}$ is the nearest point of p_a of P_a in point cloud P_b. and N_p is the normal of point p on the mesh. **F-Score (FS)**. F-Score between the two point clouds P_a and P_b given a threshold t is defined as follows:

$$F - Score(P_a, P_b, t) = \frac{2Recall \cdot Precision}{Recall + Precession} \tag{5}$$

where

$$Recall(P_a, P_b, t) = |p_a \in P_a, s.t. \min_{p_b \in P_b} d(p_a, p_b)| \tag{6}$$

$$Precision(P_a, P_b, t) = |p_b \in P_b, s.t. \min_{p_a \in P_a} d(p_b, p_a)| \tag{7}$$

We follows ONet [6], ConvONet [8] and POCO [1], we set t = 0.01.

Intersection over Union (IoU) measure the volumetric alignment between the predicted mesh and ground truth mesh. We basically sample a large number of points in unite cube of the reconstruction volume. and then count the number of points that lie in or outside of the predicted mesh and ground truth mesh. then the IOU is computed as follows:

$$IoU(M_a, M_b) = \frac{TP}{TP + FP + FN} \tag{8}$$

where TP (resp. FP, FN) are the number of the true positive points i.e. those correctly predicted as inside occupancy (reps. the number of points wrongly predicted as inside actually being outside points, and the number of points wrongly predicted as outside but actually being inside of the ground truth mesh). We sample one Million points within the reconstruction unit volume for this IOU measurement.

2 DATA AND PROCESSING

For training, we prepare three types of data for a given mesh object:
1. A given mesh is normalized to [-0.5, 0.5] before sampling. N input points will be sampled from the normalized mesh as input to the network. N were set to 10K, 3K, etc in our testing.
2. K query points will be generated by adding isotropic Gaussian noise displacement $n \sim N(0, \Sigma)$ to each sampled surface point, *i.e.* $q = p + n$, where $\Sigma \in R^{3\times3}$ is the diagonal covariance matrix with variance setting $\Sigma_{0,0} = \Sigma_{1,1} = \Sigma_{2,2} = \sigma$ defining the displacement scales. We prepare three sets of query points $K1, K2$, and $K3$, with 500,000 points in each set, and σ equals to 0.25, 0.02, 0.003, respectively for each mesh object. We then randomly pick 15%, 35%, and 50% from $K1, K2$, and $K3$, respectively, and combine them together as the final $K = 0.15\times K1 + 0.35\times K2 + 0.50\times K3$ query points for each object for training.
3. Ground truth occupancy every query point for occupancy field.

ABC, Famous, Thingi10K: We select a total of 3800 watertight meshes from ABC [5] dataset, We then split 8:2 for training and validation respectively. We use the trained model to test datasets prepared by point2surf [3], which include 100 ABC test dataset, 22 shape of Famous and 100 shapes of Thingi10K [11].

ShapeNet car: There are a total of 3094 objects that has watertight surface with no interior structures in the ShapeNet car dataset [9]. We conduct two types of evaluations, one with 3K input points, and one with 10K input points, respectively.

THuman: In order to evaluate the performance of open surface reconstruction, we use 500 human mesh of THuman2.0 [10] for human shape reconstruction. We use 7:2:1 for train, validation and test respectively. The number of input points are 10K points. We follow the same processing procedure for IFNet [2] train and test, while for POCO [1] test, we use the officially pretrained model with ShapeNet.

3 MORE QUALITATIVE RESULTS OF ABC, FAMOUS, THINGI10K

Figure 2 , Figure 3 and Figure 4 show more qualitative results of ABC [5], Famous and Thingi10K [12] respectively. POCO [1] results are obtained with officially pre-trained models.

4 MORE RESULTS OF SHAPENET CARS

Figure 5 and Figure 6 demonstrate more results of ablation studies about the network designs. Figure 5 and 6 use 3K and 10K input points respectively. Bifusion v2 refers to the network illustrated in the main manuscript with fusion module of figure 1 (a). Bifusion v0 refers to the blending weight is set to 0.5 at the output of the network. Bifusion v1 refers to using fusion module of figure 1(b).

5 MORE RESULTS ON THUMAN 2.0 HUMAN BODY SHAPE RECONSTRUCTION

Figures 7, 8, 9 and 10 list more THuman2.0 [10] results for more clear comparison among our network, baselines and POCO [1] and IFNet [2].

REFERENCES

[1] Alexandre Boulch and Renaud Marlet. 2022. Poco: Point convolution for surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6302–6314.

[2] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. 2020. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6970–6981.

[3] Philipp Erler, Paul Guerrero, Stefan Ohrhallinger, Niloy J Mitra, and Michael Wimmer. 2020. Points2surf learning implicit surfaces from point clouds. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V*. Springer, 108–124.

[4] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[5] Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. 2019. Abc: A big cad model dataset for geometric deep learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9601–9611.

[6] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4460–4470.

[7] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

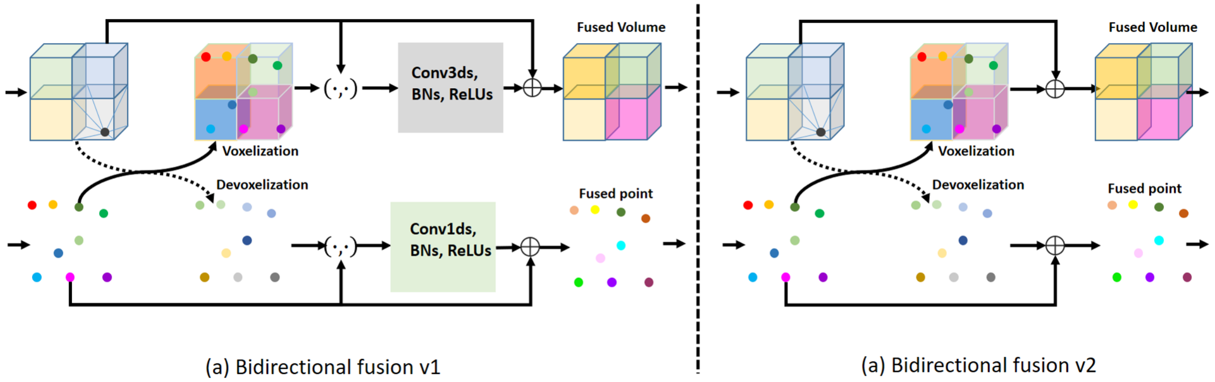


Figure 1: The two fusion modules designs for our network. (a) shows the bidirectional fusion structure version 1. (b) shows the mainly tested bidirectional fusion structure version 2.

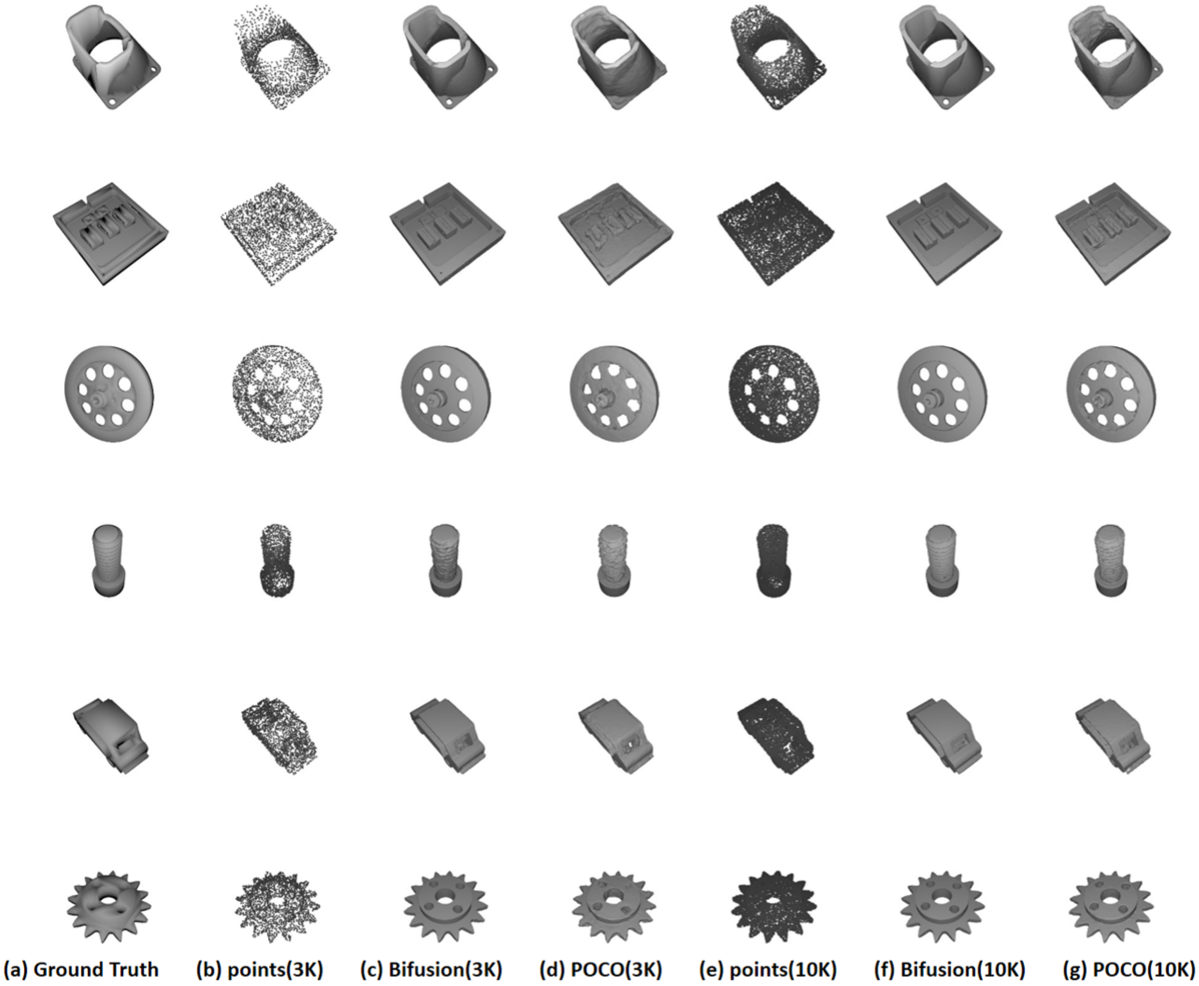
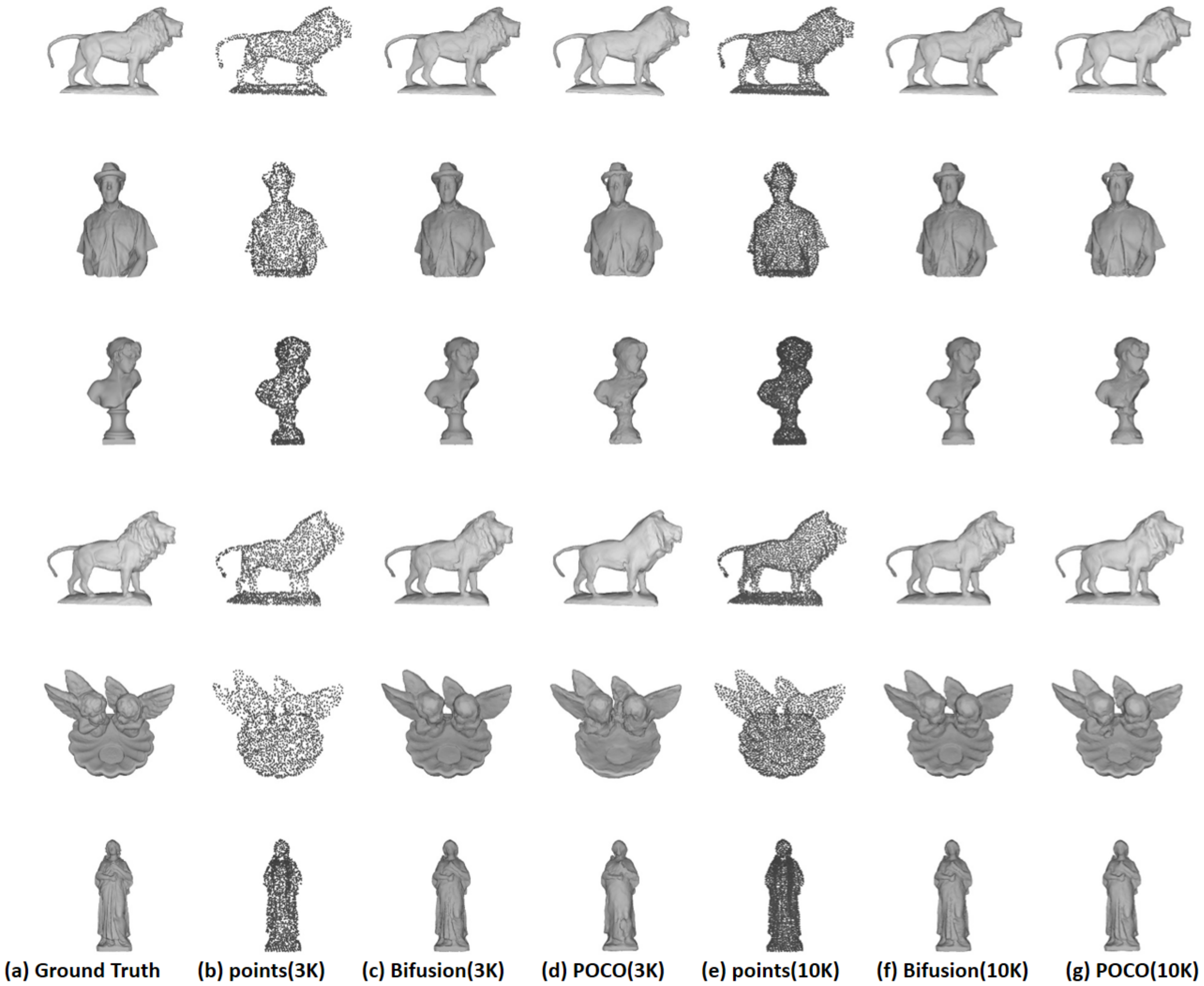


Figure 2: More ABC [5] test results. Testing data is prepared by point2surf [3]. POCO results are obtained by using officially pre-trained model.



(a) Ground Truth (b) points(3K) (c) Bifusion(3K) (d) POCO(3K) (e) points(10K) (f) Bifusion(10K) (g) POCO(10K)

Figure 3: More Famous dataset test results. Testing data is prepared by point2surf [3]. POCO results are obtained by using officially pre-trained model.

[8] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. 2020. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 523–540.

[9] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. 2019. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in neural information processing systems* 32 (2019).

[10] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. 2021. Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*.

[11] Qingnan Zhou and Alec Jacobson. 2016. Thingi10K: A Dataset of 10, 000 3D-Printing Models. *ArXiv* abs/1605.04797 (2016). https://api.semanticscholar.org/CorpusID:39867743

[12] Qingnan Zhou and Alec Jacobson. 2016. Thingi10K: A Dataset of 10,000 3D-Printing Models. *arXiv preprint arXiv:1605.04797* (2016).

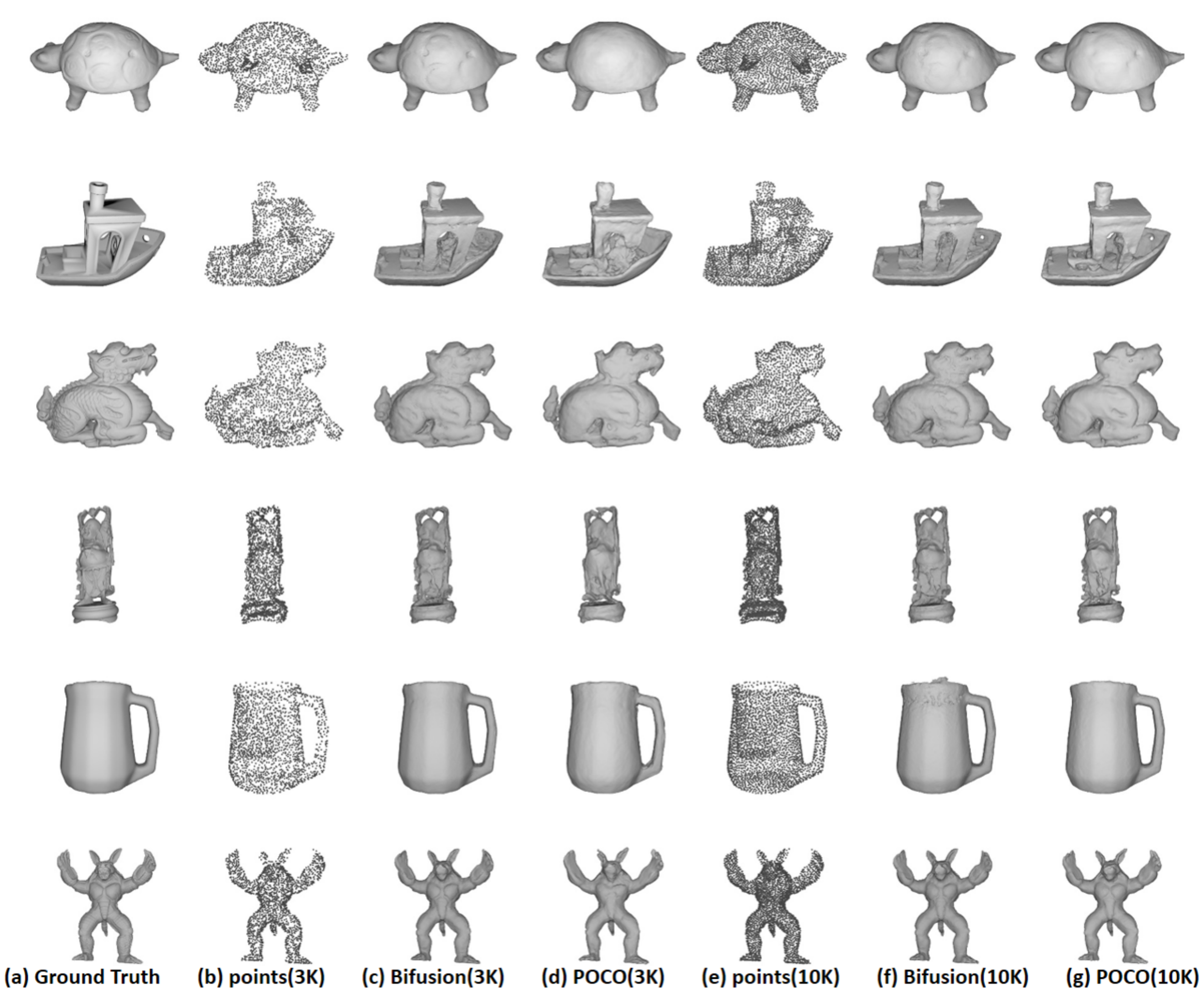


Figure 4: More results on Thingi10K. Testing data is prepared by point2surf [3]. POCO results are obtained by using officially pre-trained model.

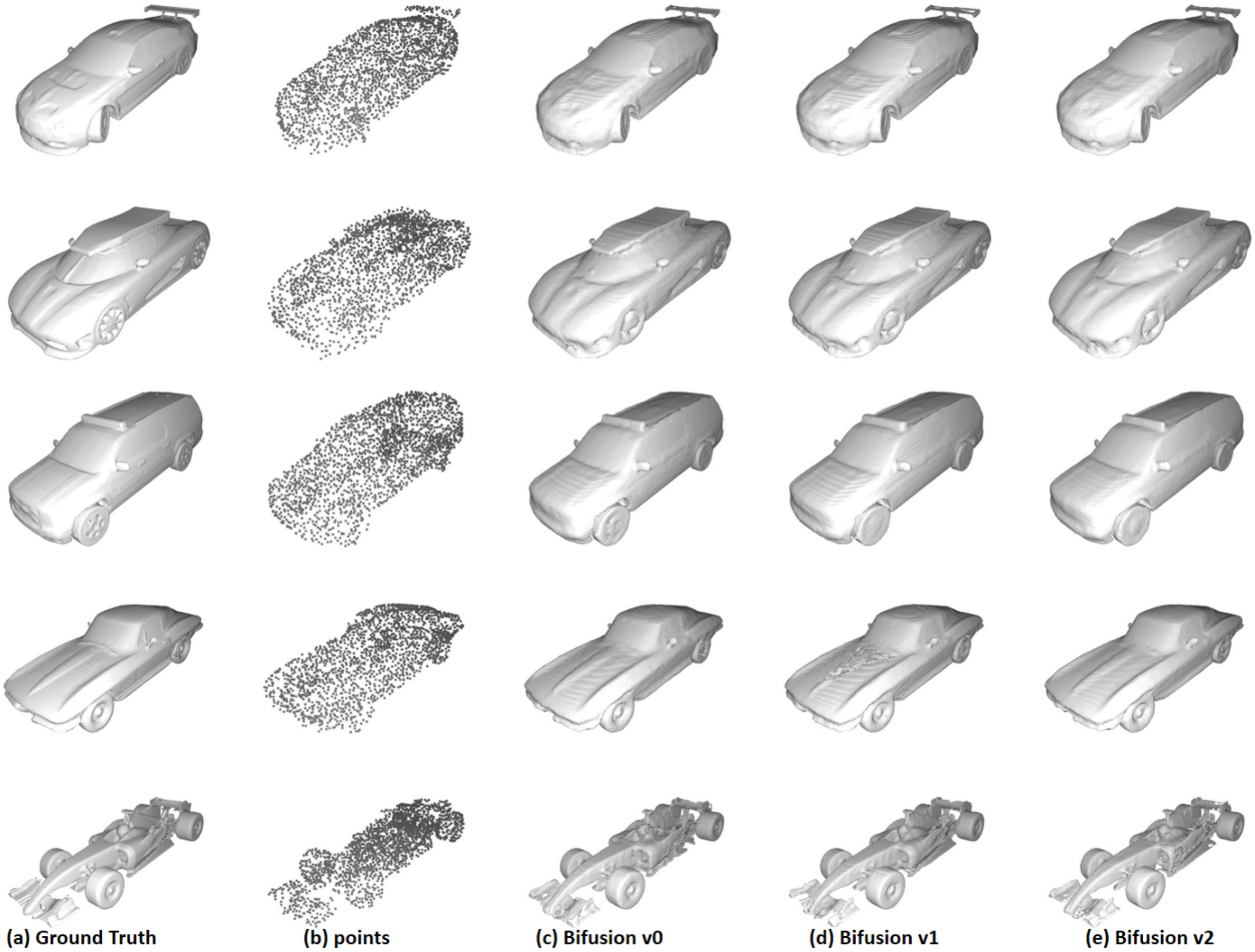


Figure 5: Ablation study with various network designs with 3K input points. Bifusion v0 uses simple average of occupancy O1 of volume branch and occupancy O2 of point branch. i.e. *omega* = 0.5. Bifusion v1 use complex fusion module instead of the version shown in figure of fusion block. Please refer to the supplementary manuscript for detailed structure design. Bifusion v2 refers to the main network demonstrated in this manuscript of figure of the network which uses fusion modules of figure of fusion block.

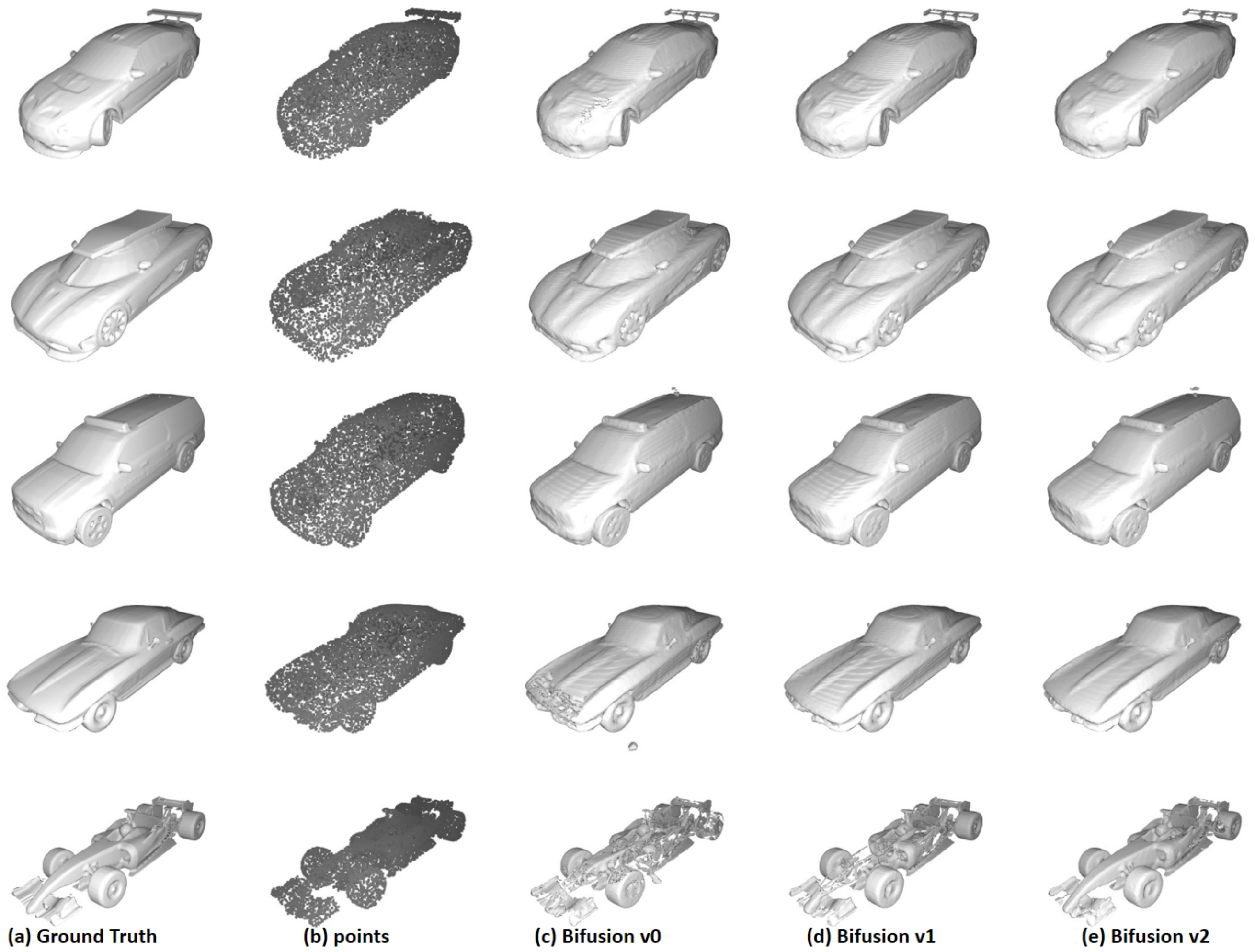


Figure 6: Ablation study with various network designs with 10K input points. Bifusion v0 use simple average of occupancy O1 of volume branch and occupancy O2 of point branch. i.e. *omega* **= 0.5. Bifusion v1 use complex fusion module instead of the version shown in figure of fusion block. Please refer to the supplementary manuscript for detailed structure design. Bifusion v2 refers to the main network demonstrated in this manuscript of figure of network which uses fusion modules of figure of fusion block.**

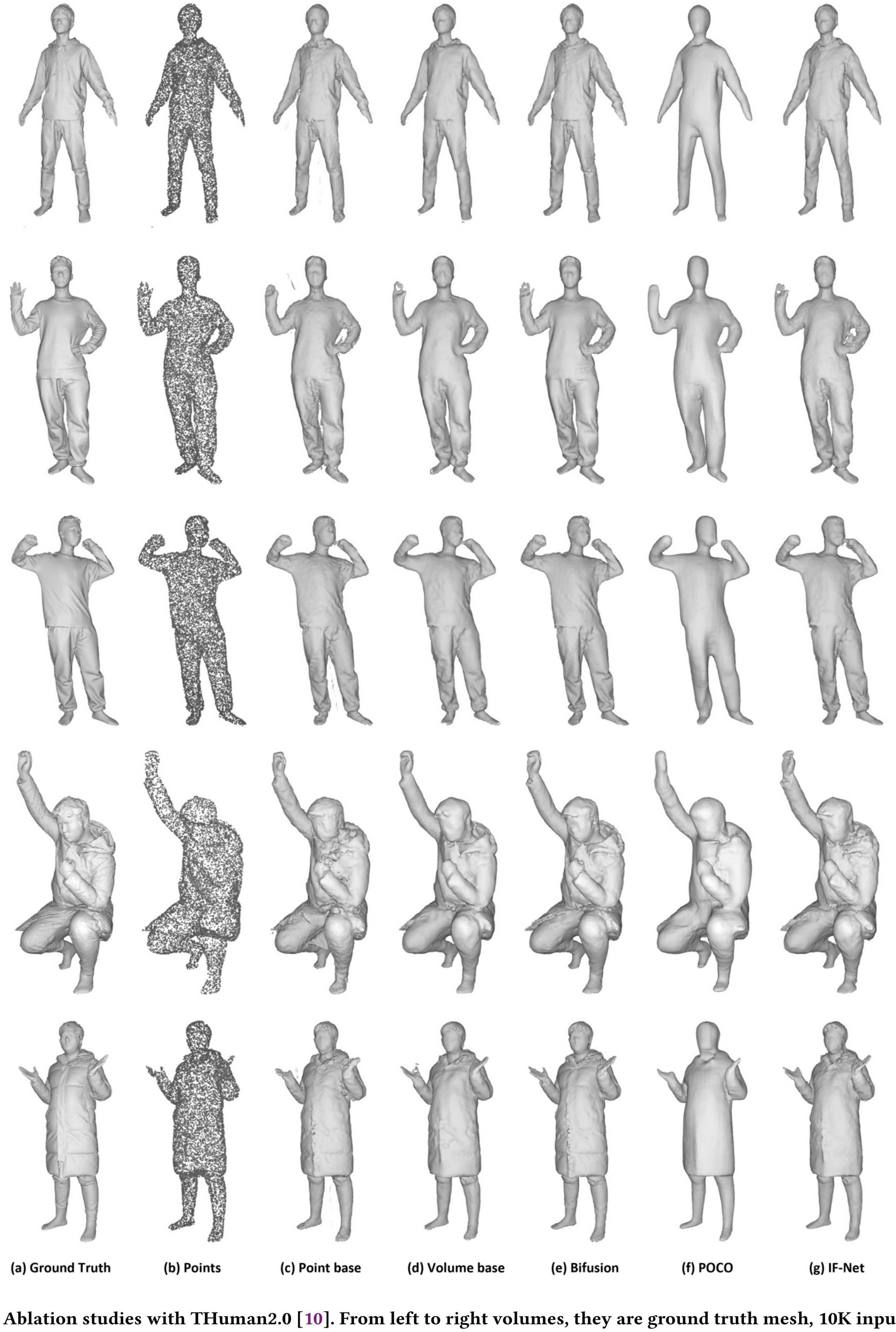


Figure 7: Ablation studies with THuman2.0 [10]. From left to right volumes, they are ground truth mesh, 10K input points, point base network results, volume base network results, Bifusion v2 results, POCO [1] and IFNet [2] results respectively.

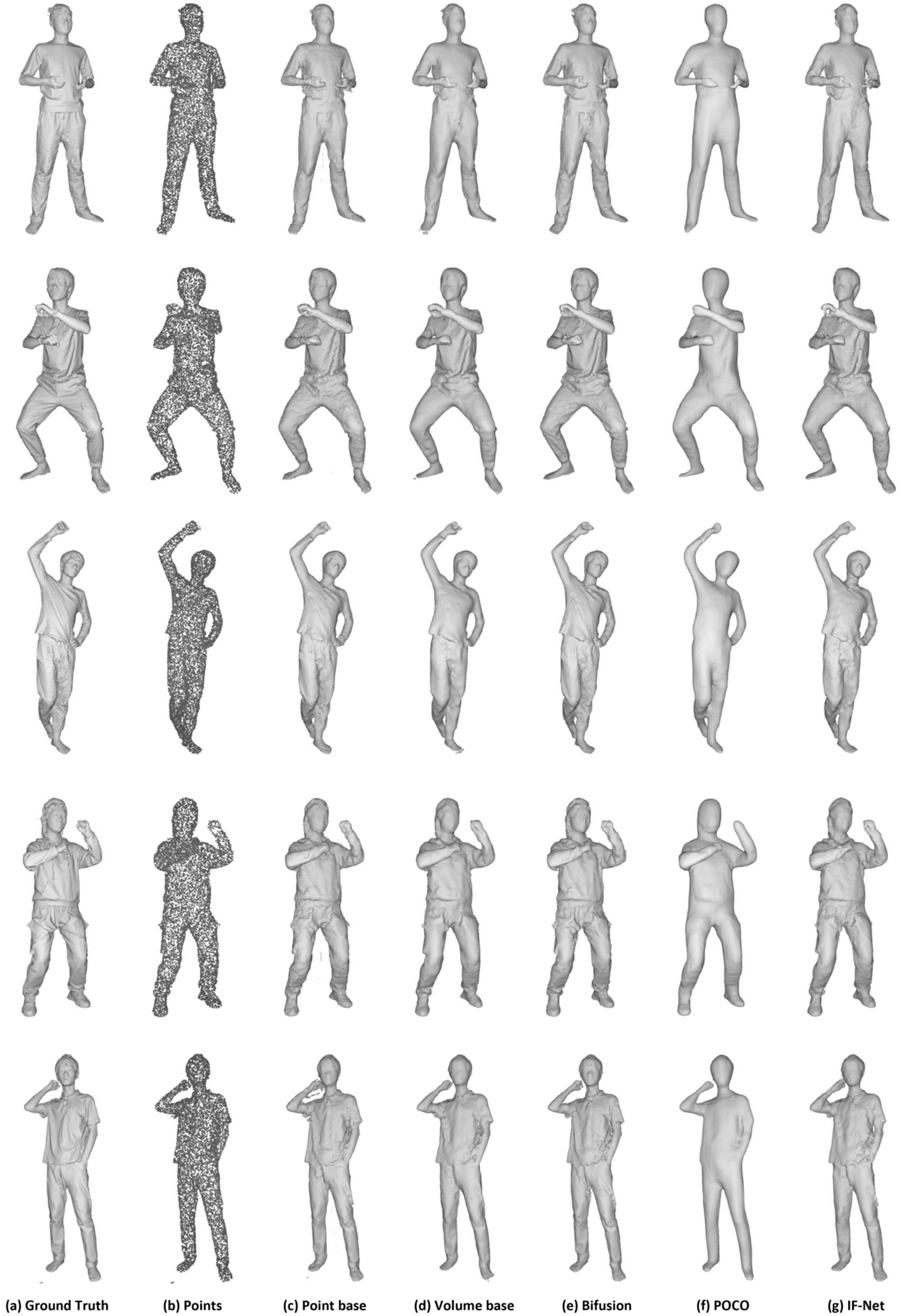


Figure 8: Figure 7 continued

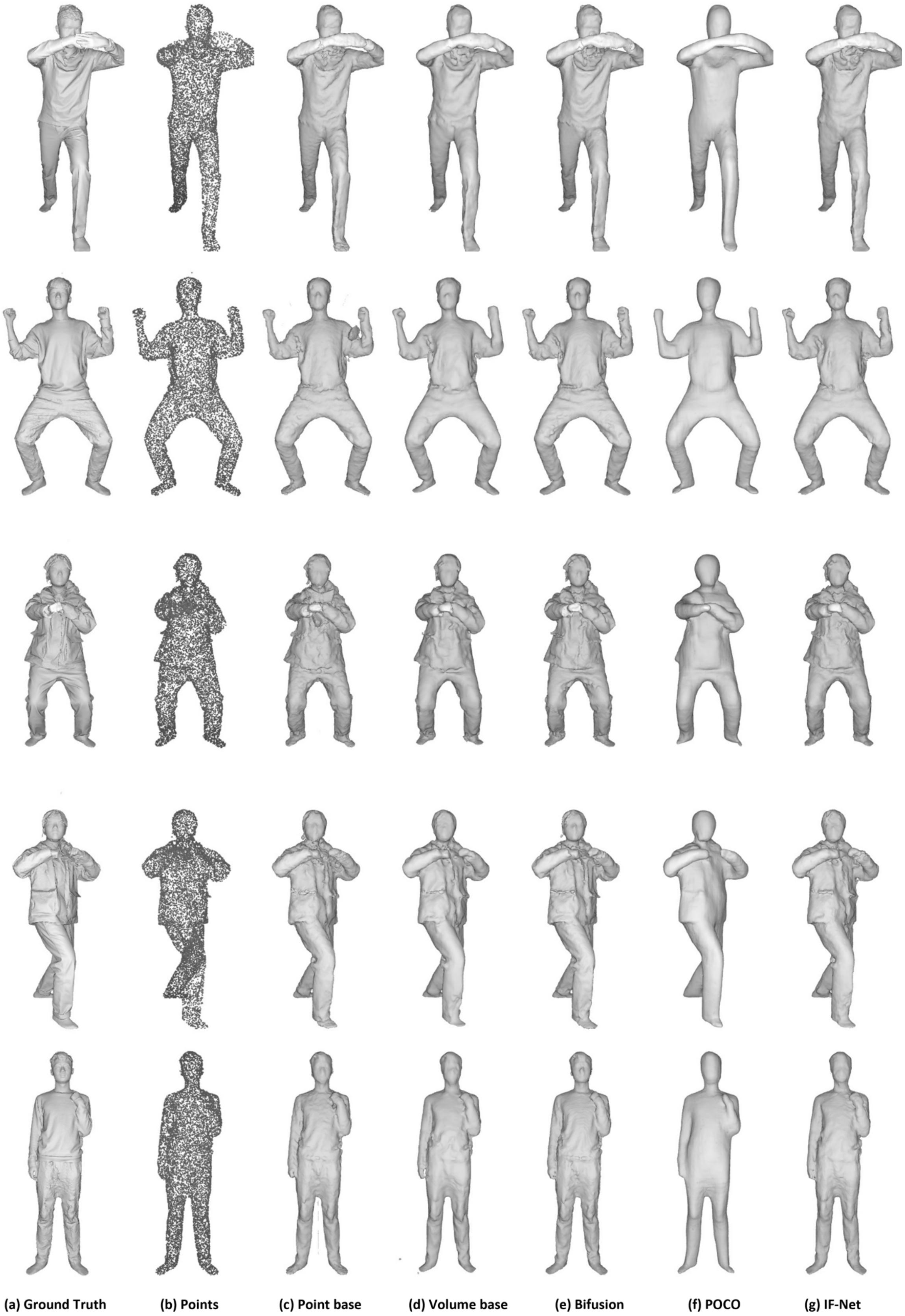


Figure 9: Figure 7 Continued

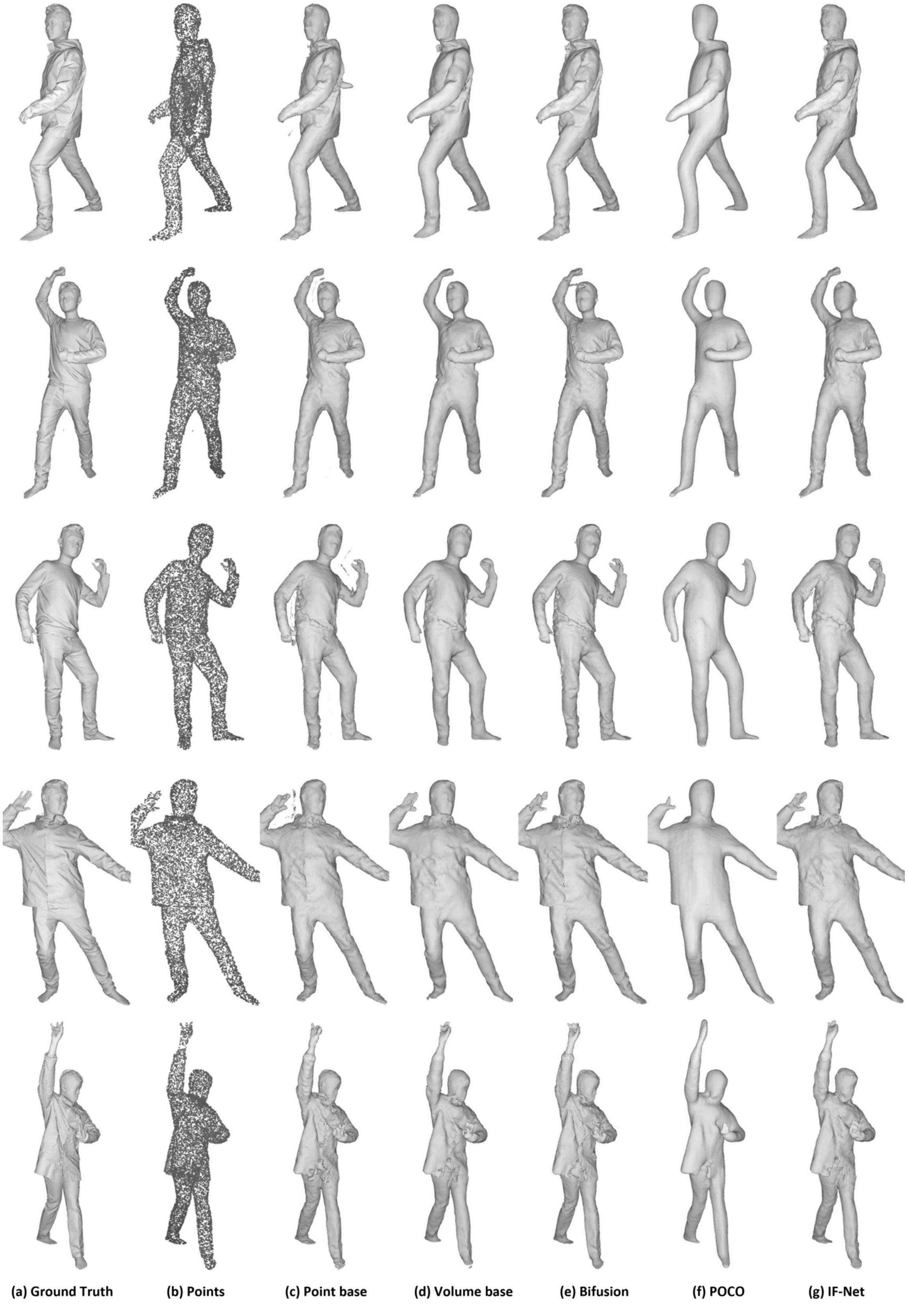


Figure 10: Figure 7 Continued