Supplementary Material for Learning Hand-Held Object Reconstruction from In-The-Wild Videos

Aditya Prakash Matthew Chang Matthew Jin Saurabh Gupta
University of Illinois Urbana-Champaign
https://ap229997.github.io/projects/horse

In this supplementary document, we first provide additional implementation details about the multiview mask supervision and learned shape priors used in our architecture (Sec. 1). Next, we provide examples of hand-object tracks on VISOR (Fig. 1, Fig. 2), qualitative comparisons of our HORSE model with AC-SDF on MOW (Fig. 3, Fig. 4) & HO3D (Fig. 5), failure cases of our model (Fig. 6) and visualizations of cross-section maps used in the discriminator (Fig. 7). Moreover, the supplementary video contains 3D visualizations of shape predictions from HORSE and AC-SDF and provides a summary of our key ideas & contributions.

1. Architecture Details

Supervision from Masks in Multiple Views: We supervise the occupancy network f via consistency in masks in multiple views of the object. Let's assume we have multiple views $\{I_1, \ldots, I_n\}$ and corresponding segmentation masks for objects of interaction in these n views $\{M_1, \ldots, M_n\}$. Given this setup, we compute the occupancy labels for a point $\mathbf{x}$ as $\mathbf{s}^{gt} = \cap_{i=1}^{n} M_i(\mathbf{x}_{p_i})$ and train using $\mathcal{L}_{\text{visual-hull}} = \text{CE}(\mathbf{s}, \mathbf{s}^{gt})$ where CE is cross-entropy loss, $\mathbf{s} = f(\mathbf{x}; I, \theta^a, \theta^w, K)$ is the predicted occupancy at $\mathbf{x}$ and $\mathbf{x}_{p_i} = \pi_{\theta_i^w}(\mathbf{x})$ is the projection of $\mathbf{x}$ into the image I_i using FrankMocap [6] predictions of the hand pose and camera parameters.

An important consideration is assigning ground truth labels to the 8192 sampled points since we rely on FrankMocap predicted hand pose and camera parameters which may be inaccurate. First, we uniformly sample 4096 3D points in [-1,1]^3 in the wrist coordinate frame and consider the points projecting into the object mask in all views as occupied. Since all these 4096 points may not be occupied, we use rejection sampling to repeat the procedure, for maximum of 50 times, until we get 4096 occupied points. Moreover, due to hand occlusions and errors in FrankMocap predictions, it is possible that some 3D points belonging to the object are not projected into the object masks but we do not want the network to predict these points as unoccupied. So, we adopt additional strategies for sampling unoccupied points. We disregard points which project onto the object mask in some views and hand mask in other views as these points could belong to object due to hand occlusion. Also, all points projecting into the hand mask in all views and vertices of the MANO [5] mesh are labeled as unoccupied.

Hand-held Object Shape Prior Supervision: We adopt an adversarial training framework [2] to jointly build a 3D shape prior using synthetic ObMan objects while providing the supervision for training f. We train a discriminator g to differentiate between real (synthetic shapes from ObMan dataset) and generated shapes (shapes predicted by f). We derive supervision for f by encouraging it to predict shapes that are real as per the discriminator g.

Building this prior in 3D is computationally challenging due to large number of quieries required to the occupancy network. Instead, we build this prior on occupancy values in arbitrary 2D slices (that pass through the origin in the wrist coorindate frame) of real and predicted 3D shapes. This involves sampling the function f on a 2D grid, of dimension 32x32, placed on a random plane passing through the origin.

We implement the discriminator as a 3-layer convolutional network which takes as input a 32x32 slice and outputs a 512 dimensional feature vector, followed by a 1-layer MLP with 128 hidden nodes. We use 64, 128 & 256 filters in the 1^{st}, 2^{nd} & 3^{rd} convolutional layers respectively and use filter size of 5x5 for all convolutional layers. We use ReLU activation function for all layers except the last layer and use BatchNorm after each convolution.

Supervision for f is obtained by computing gradients through g on the sampled occupancy values so as to maximize the realism of the occupancy slices output by f. Following [3], we use the least squares formulation [4] for training the discriminator. The shape prior loss takes the form $\mathcal{L}_{\text{shape-prior}} = g(\text{slice}(f(\cdot; I, \theta^a, \theta^w, K)))^2$, where slice function samples a 2D slice using the occupancy function f and g is the discriminator that has been trained to predict 0 for real samples and 1 for generated samples using L2 loss, $\mathcal{L}_{\text{disc}}$. The training is done in an alternation fashion where g is updated twice for every update of f, which is trained using $\mathcal{L}_{\text{visual-hull}}$, $\mathcal{L}_{\text{consistency}}$, & $\mathcal{L}_{\text{shape-prior}}$. We use Adam optimizer

with learning rate of $1e{-}5$ for both f and g. We set loss weights as $\lambda_{\text{shape-prior}} = 0.25$, $\lambda_{\text{disc}} = 0.25$.

2. Visualizations

Hand and object tracks on VISOR: Using existing hand pose estimation techniques [6], we are able to track the objects in relation to hands through time, in in-the-wild videos. We visualize these tracks along with object masks from the VISOR dataset [1] in Fig. 1. This form of data, where objects move rigidly relative to hands, is used to train our proposed formulation to learn 3D geometry of hand-held objects from in-the-wild videos. However, we rely on FrankMocap predictions, which may not be always accurate throughout the video clip, as shown in Fig. 2. We automatically filter these sequences using uncertainty estimates from the procedure described in Sec. 4 of the main paper.

Qualitative comparison of HORSE and AC-SDF: We compare the mesh generated by our model and AC-SDF trained on ObMan on object generalization splits of MOW (Fig. 3, Fig. 4) & HO3D (Fig. 5) datasets. For this, we sample points uniformly in a 64x64x64 volume, predict their occupancies or SDF from the network and run marching cubes to get the mesh. We project the mesh into the input image and render it in different views. We observe that our model is able to capture the visual hull of the object, as evidenced by the projection of the mesh onto the image, and can generate more coherent shapes than AC-SDF, which often reconstructs disconnected and scattered shapes.

Failure cases of HORSE: From visualizations in Fig. 6. We observe that our model sometimes generates enlarged or protruded shapes. This could be due to inaccurate hand pose during training because of which the model gets confused about which points are inside or outside. Moreover, some objects are only partially visible in the cropped image, which is input to the model, due to which the pixel-aligned features may not represent the object semantics well.

Cross-section maps for discriminator: We visualize the 2D cross-sections from the intersection of the sampled planes with the 3D shape of the object in Fig. 7. Each of these cross-section map is 32x32 dimensional and we show 120 (8x15 grid) maps in each image, with the red region denoting pixels inside the object and yellow region representing pixels outside. These cross-sections are fed as input to the discriminator which is trained to distinguish between the cross-sections from our model and cross-sections from the synthetic shapes in the ObMan dataset. We show the cross-sections from our predictions on the MOW dataset without the discriminator and with the discriminator. We observe that holes in the cross-section maps and size of the blobs get reduced after training with discriminator.

References

[1] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *NeurIPS Track on Datasets and Benchmarks*, 2022. 2, 3

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 1

[3] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[4] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[5] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 2017. 1

[6] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3D hand and body motion capture by regression and integration. *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2021. 1, 2, 3

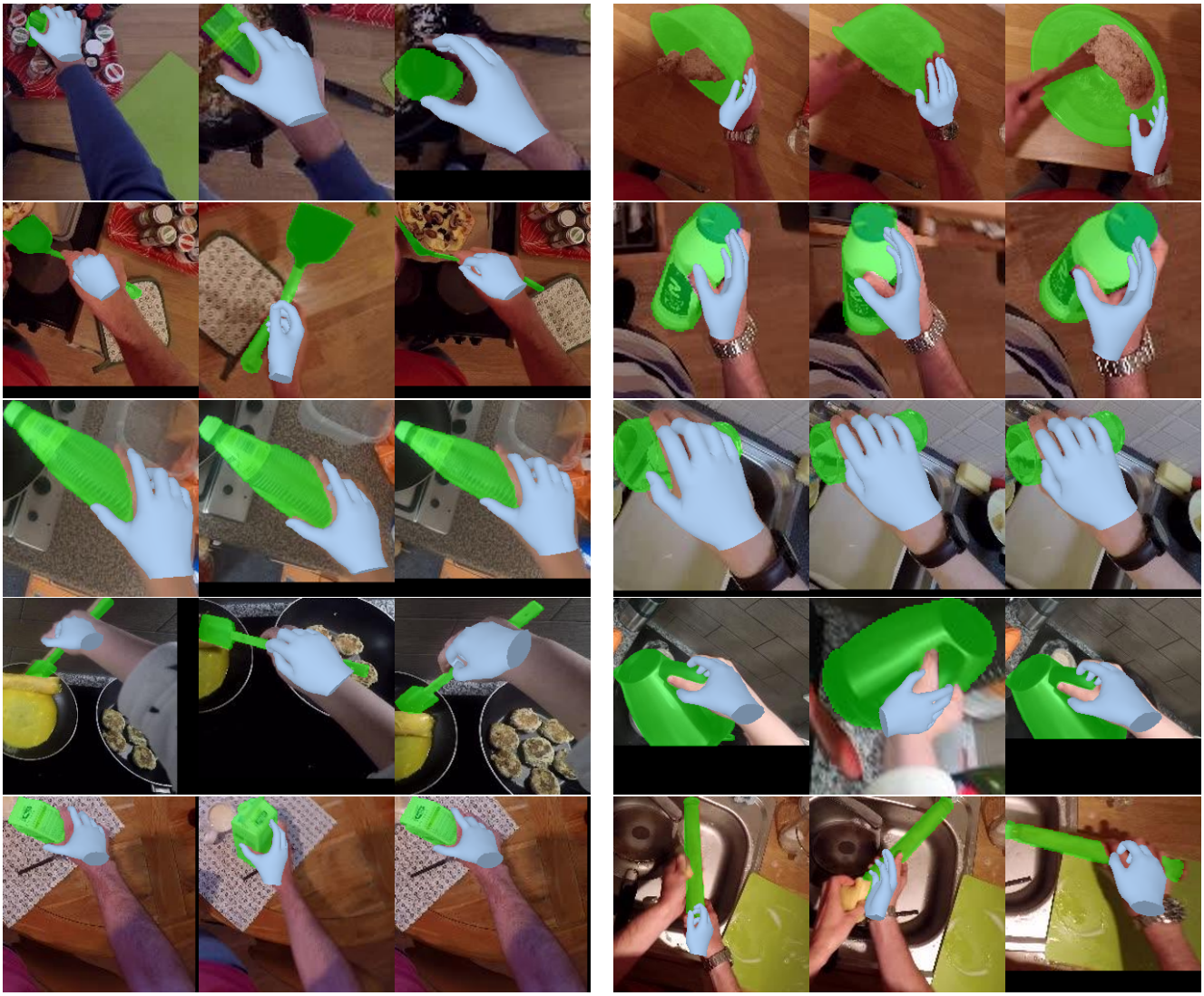

Figure 1: Hand-object tracks on VISOR. We visualize the hand predictions from FrankMocap [6] and the object masks from VISOR [1] on several video clips. We observe that FrankMocap predictions are reaonsably accurate for most clips involving several objects.

Figure 2: Inaccurate hand pose. We observe several cases where the predicted hand pose from FrankMocap in some frames is not accurate.

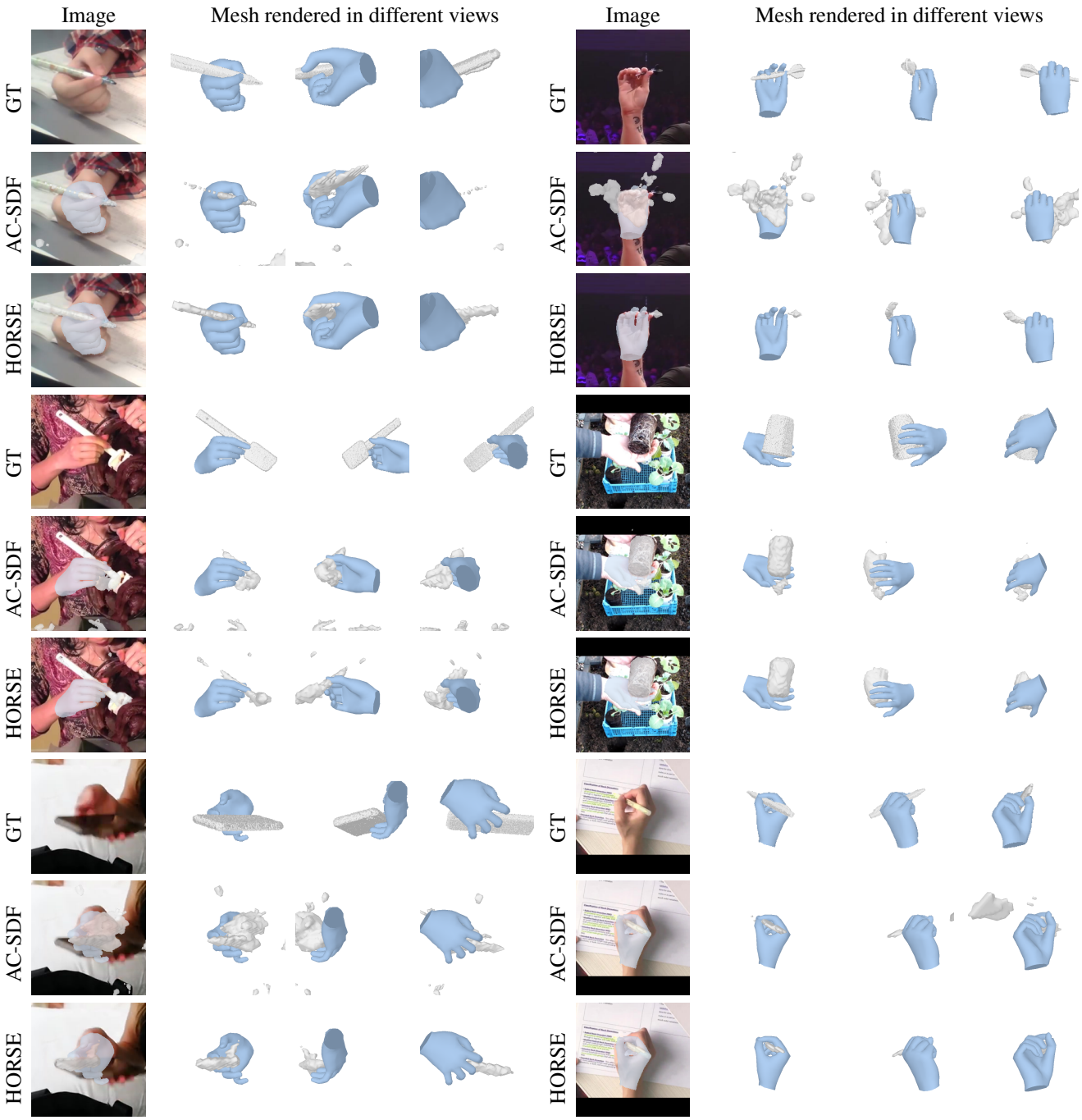

Figure 3: Mesh visualizations on MOW object generalization split. We show the object mesh projected onto the image and rendered in different views for our HORSE model and compare with the AC-SDF model trained on ObMan dataset with 3D supervision (best baseline model). We also show the ground truth (GT) object model. We observe that our model is able to predict the object shape more accurately than AC-SDF which often reconstructs smaller and disconnected shapes.

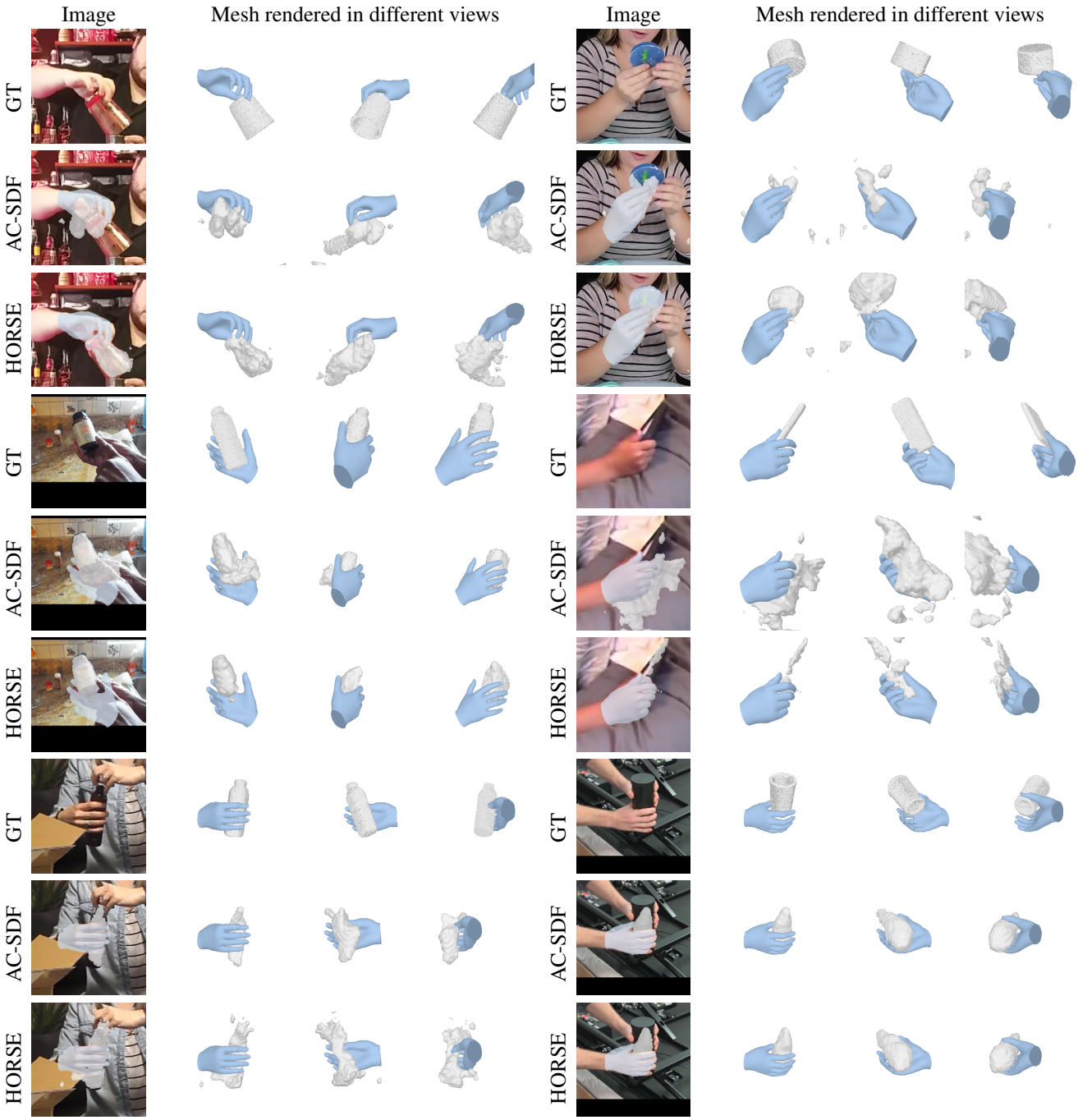

Figure 4: Mesh visualizations on MOW object generalization split. We show the object mesh projected onto the image and rendered in different views for our HORSE model and compare with the AC-SDF model trained on ObMan dataset with 3D supervision (best baseline model). We also show the ground truth (GT) object model. We observe that our model is able to predict the object shape more accurately than AC-SDF which often reconstructs smaller and disconnected shapes.

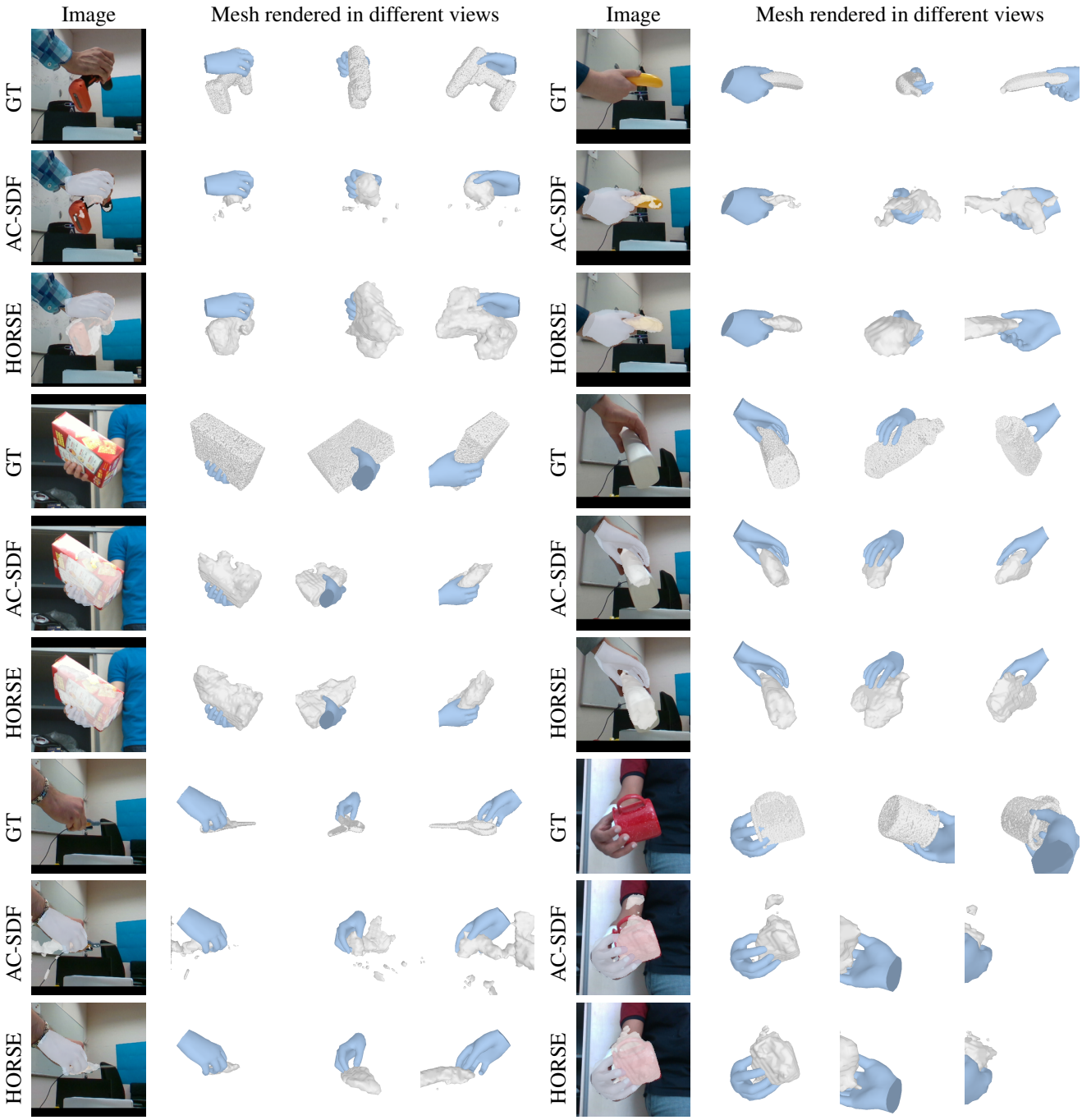

Figure 5: Mesh visualizations on HO3D object generalization split. We show the object mesh projected onto the image and rendered in different views for our HORSE model and compare with the AC-SDF model trained on ObMan dataset with 3D supervision (best baseline model). We also show the ground truth (GT) object model. We observe that our model is able to predict the object shape more accurately than AC-SDF which often reconstructs smaller and disconnected shapes.

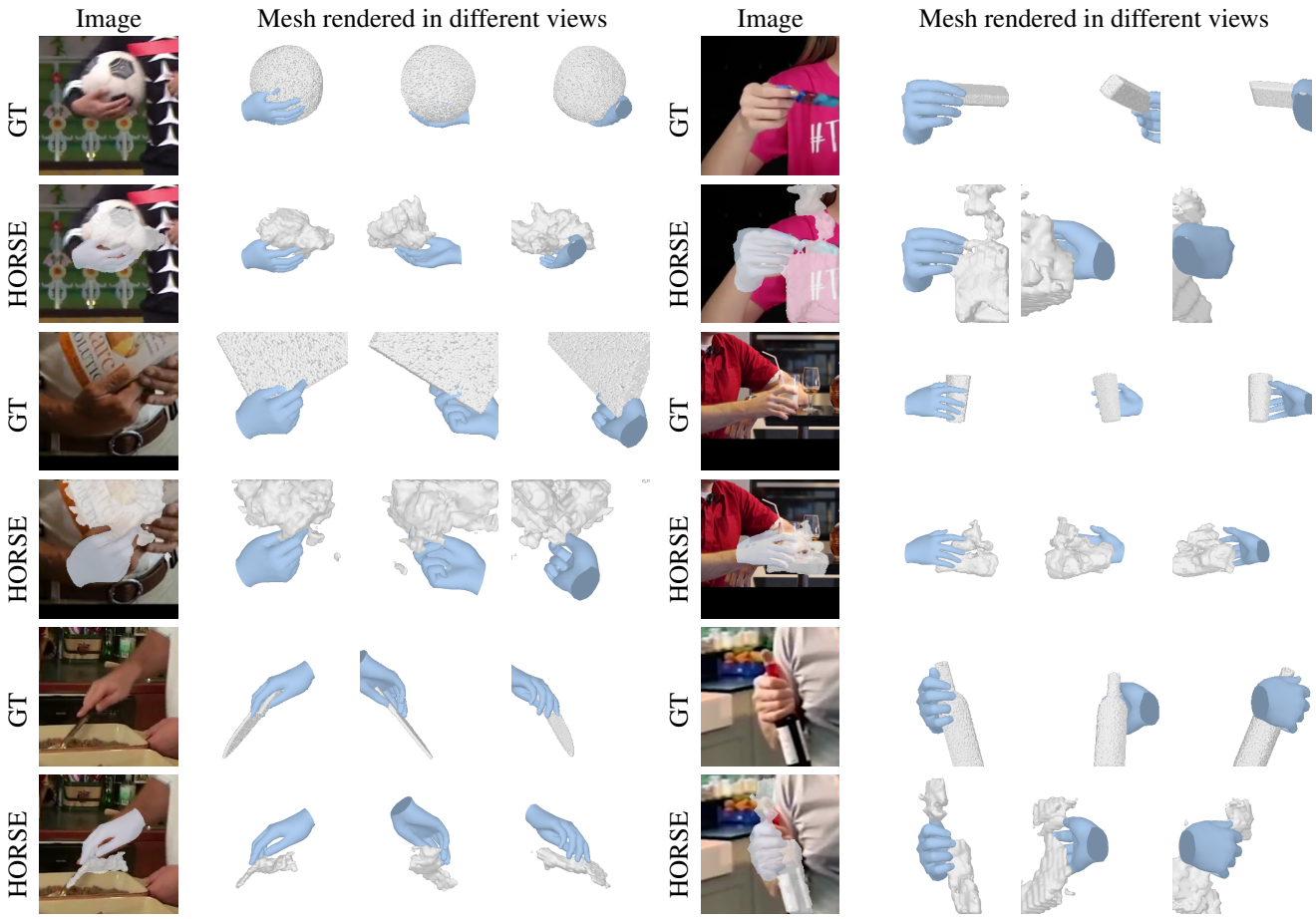

Figure 6: Failure cases of HORSE on MOW object generalization split. We visualize some failure cases of our HORSE model MOW dataset. We observe that our methods sometimes generates enlarged or protruded shapes. This could be due to inaccurate hand pose during training because of which the model gets confused about which points are inside or outside. Moreover, some objects are only partially visible in the cropped image, which is input to the model, due to which the pixel-aligned features may not represent the object semantics well.

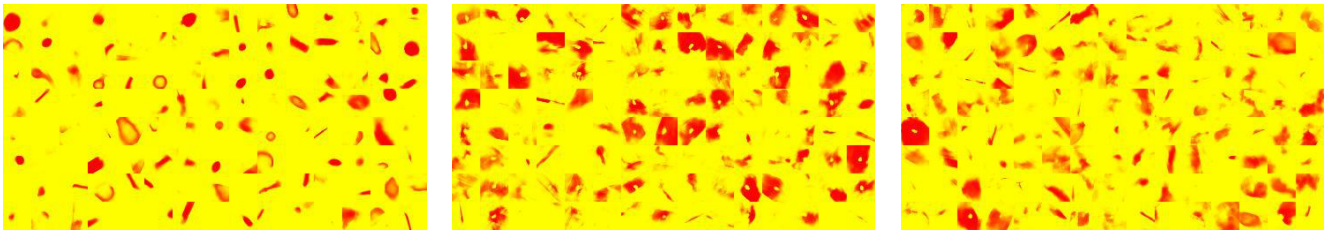

Figure 7: Visualizations of cross-section maps. We show the 2D cross-sections from the intersection of the sampled planes with the 3D shape of the object. Each of these cross-section map is 32x32 dimensional and we visualize 120 (8x15 grid) maps in each image, with the red region denoting pixels inside the object and yellow region representing pixels outside. These cross-sections are fed as input to the discriminator which is trained to distinguish between the cross-sections from our model and cross-sections from the synthetic shapes in the ObMan dataset (left). We show the cross-sections from our predictions on the MOW dataset without the discriminator (middle) and with the discriminator (right). We observe that holes in the cross-section maps and size of the blobs get reduced after training with discriminator.