arXiv:2107.00887v1 [cs.CV] 2 Jul 2021

HO-3D_v3: Improving the Accuracy of Hand-Object Annotations of the HO-3D Dataset

Shreyas Hampali[(1)], Sayan Deb Sarkar[(1)], Vincent Lepetit[(2,1)]
[(1)]Institute for Computer Graphics and Vision, Graz University of Technology, Graz, Austria
[(2)]Université Paris-Est, École des Ponts ParisTech, Paris, France
{<firstname>.<lastname>}@icg.tugraz.at
Project page: https://www.tugraz.at/index.php?id=40231

1 What's new in the 'v3' version of HO-3D?

HO-3D is a dataset providing image sequences of various hand-object interaction scenarios annotated with the 3D pose of the hand and the object and was introduced in [1] as version HO-3D_v2. The annotations were obtained automatically using an optimization method introduced in the original paper.

HO-3D_v3 provides **more accurate** annotations for both the hand and object poses thus resulting in **better estimates of contact regions** between the hand and the object. The new annotations were obtained using an improved version of the optimization technique of [1] as detailed in Section 2 below. Table 1 shows the statistics of the HO-3D_v2 and HO-3D_v3 versions of the dataset.

The improvement in accuracy of HO-3D_v3 annotations is shown in Table 2 based on 50 manually annotated frames. Figure 1 shows the contact regions between the hand and the objects over the hand surface averaged over the entire dataset together with the mean penetration distance. HO-3D_v3 annotations clearly results in more contacts compared to HO-3D_v2.

	Train set #images	Test Set #images	#objects	#subjects	Release date
HO-3D_v2	66,034	11,524	10	10	9 Jan 2020
HO-3D_v3	83,325	20,137	10	10	1 Jul 2021

Table 1: Statistics of the HO-3D_v2 and HO-3D_v3 datasets.

2 Algorithmic improvements

We made the following three changes to our method proposed in [1] in order to improve the accuracy of the pose estimates.

	mean (mm)	std (mm)
HO-3D_v2	12.31	3.8
HO-3D_v3	8.12	2.2

Table 2: Accuracy of our estimations with respect to manual annotations of fingertips using the point cloud built from all camera depth maps. We manually annotated 50 frames with all 5 fingertip locations. The frames were specifically chosen from parts of the sequence where HO-3D_v2 provides poor estimates.

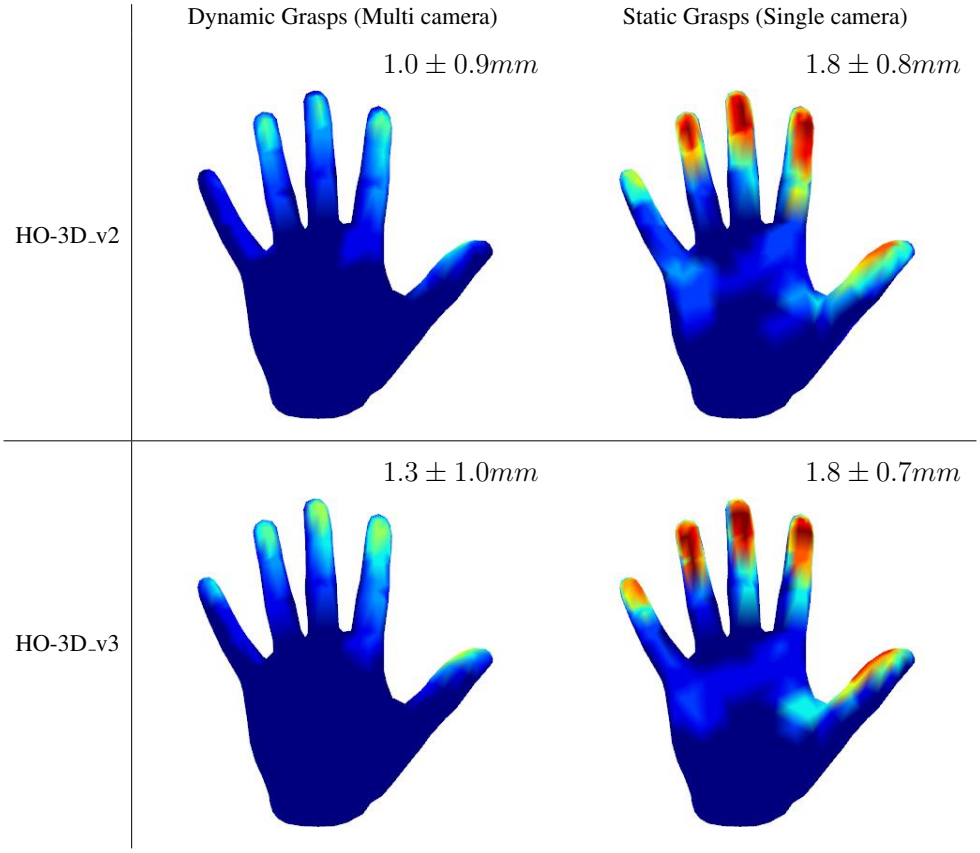


Figure 1: Contact maps for the HO-3D_v2 and HO-3D_v3 versions of the datasets for dynamic and static grasps. 'Red' regions denote complete contact and 'Blue' regions denote no contact. The mean penetration between hand and object are provided in the inset. The 'v3' dataset shows more contact regions than the 'v2' dataset with significant improvement in the case of dynamic grasps.

2.1 Cross-entropy based Silhouette Discrepancy Term

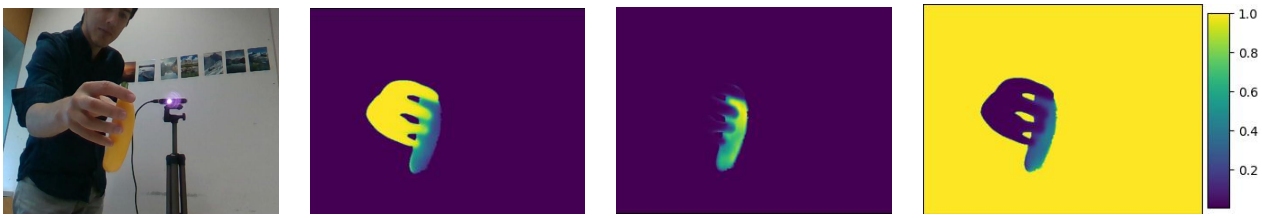
We replace the L2 term for the Silhouette error in Eq. (4) of [1] with the cross-entropy loss. Moreover, instead of obtaining the segmentations for the hand and object from a network trained on synthetic data as in [1], we use the segmentation prediction by MSeg [2].

MSeg predicts a segmentation with 194 categories. We remap these categories to 3 classes: *ob-*

ject, *person* and *background*. Figure 2 shows the confidence maps predicted for these 3 categories using the MSeg network on an example.

Denoting the segmentation confidence maps for camera v by $S_v(c)$, where $c \in \{$*object, person, background*$\}$ and the rendered segmentation by $S_v^R(c)$ (*i.e.* a pixel in $S_v^R(c)$ has value 1 if lying on class c and 0 otherwise), the silhouette term is defined as

$$E_{\text{mask}} = \sum_c S_v(c) \cdot S_v^R(c) \; . \tag{1}$$



Color image | Hand confidence map | Object confidence map | Background confidence map

Figure 2: Confidence maps for the hand, object, and background classes obtained by merging the predictions made by MSeg [2] for 194 categories.

2.2 Anatomically Oriented Joint Axes

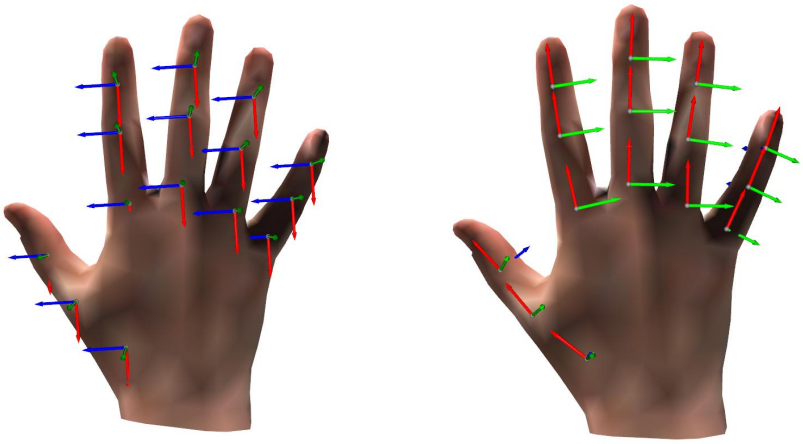
We optimize directly on the joint angles instead of its PCA components. We have found this strategy helpful to 1) reach grasp poses that were not reachable when using PCA components and 2) avoid getting stuck in local minima especially when optimizing on segmentation and depth maps. The joint axes in the MANO [3] model are not aligned along the anatomical directions of fingers which makes the task of defining the joint angle limits difficult. Figure 3a shows the joint axes in the MANO model and Figure 3c shows the resulting implausible pose due to inaccurate joint angle limits. Figure 3b shows the manually aligned joint axes along the anatomical direction and the resulting pose in the HO-3D_v3 dataset.

2.3 Improved Repulsion Term

We replace the physical plausibility term E_{phy} in Eq. (9) of [1] with a simpler sphere-based repulsion term. More specifically, we obtained a "sphere-based representation" for the hand and the object by populating their inner volumes with spheres as shown in Figure 4. We define the repulsion term between the ith hand sphere and the jth object sphere with radii denoted by r_i^h, r_j^o and centers denoted by c_i^h, c_j^o, respectively as:

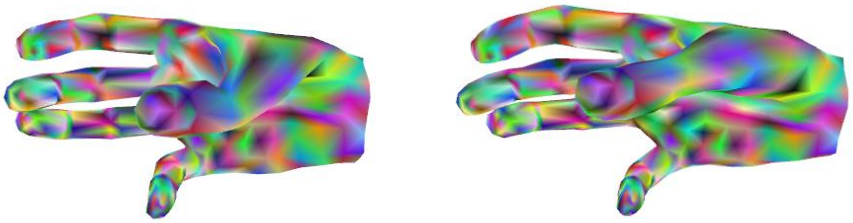
$$E_{\text{phy}} = \sum_i \sum_j \max(0, r_i^h + r_j^o - ||c_i^h - c_j^o||_2 - t) \; , \tag{2}$$

where t is the allowed penetration between the hand and the object. We use $t = 2mm$.



(a) Joint axes not aligned along anatomical direction in MANO [3]

(b) Joint axes manually aligned with joint direction in our method



(c) Twisted thumb when using non-anatomically aligned joint axes

(d) More natural pose when using anatomically aligned joint axes

Figure 3: We align the joint axes along anatomical direction of the fingers resulting in more natural poses in the HO-3D_v3 dataset.

3 Evaluations

We report here the improvement in hand pose accuracy of HO-3D_v3 with respect to HO-3D_v2. We manually annotated the point cloud of 50 frames with finger tip locations to measure the accuracy of our method. The point cloud is obtained by combining the point cloud from all cameras in the multi-camera setup as shown in Figure 5. Further, the 50 frames were specifically chosen from parts of the HO-3D_v2 sequences where the hand pose accuracy is poor. As shown in Table 2, HO-3D_v3 shows 4mm improvement in accuracy compared to HO-3D_v2.

Figure 2 shows the contact maps for dynamic and static grasp sequences for HO-3D_v3 and HO-3D_v2 datasets along with mean penetration between hand and object. We consider a point on the surface of the hand to be in contact with the object if the distance to the object suface is less than 4mm. We allow this tolerance to account for depth errors and error in camera calibration. The poses in HO-3D_v3 exhibit more hand-object contact while having similar inter-penetration. The contact maps for each of the sequences in the dataset are shown in Figures 6, 7 and 8.

(a) *Bleach Cleanser* object with its inner volume populated with spheres

(b) Hand model and its sphere representation

Figure 4: Sphere-based object and hand models.

Figure 5: Point cloud from all cameras along with manual annotation of the finger tips.

References

[1] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3

[2] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. MSeg: A composite dataset for multi-domain semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[3] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):245:1–245:17, November 2017. 3, 4

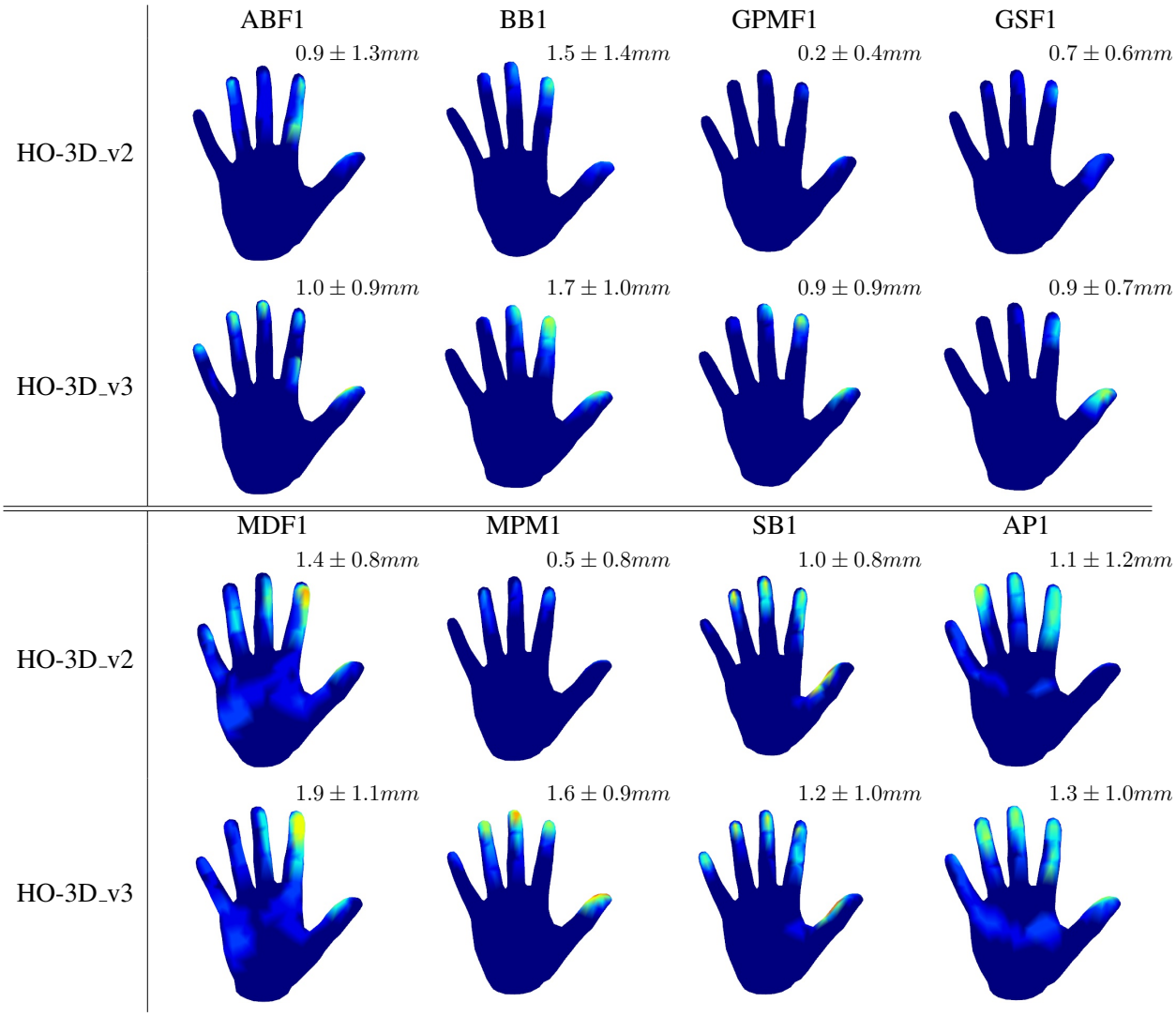


Figure 6: Comparison of contact maps between the 'v2' and 'v3' versions of the HO-3D dataset for all the dynamic grasp sequences. The mean penetration for each sequence is provided in the inset. We see an overall improvement in grasp quality.

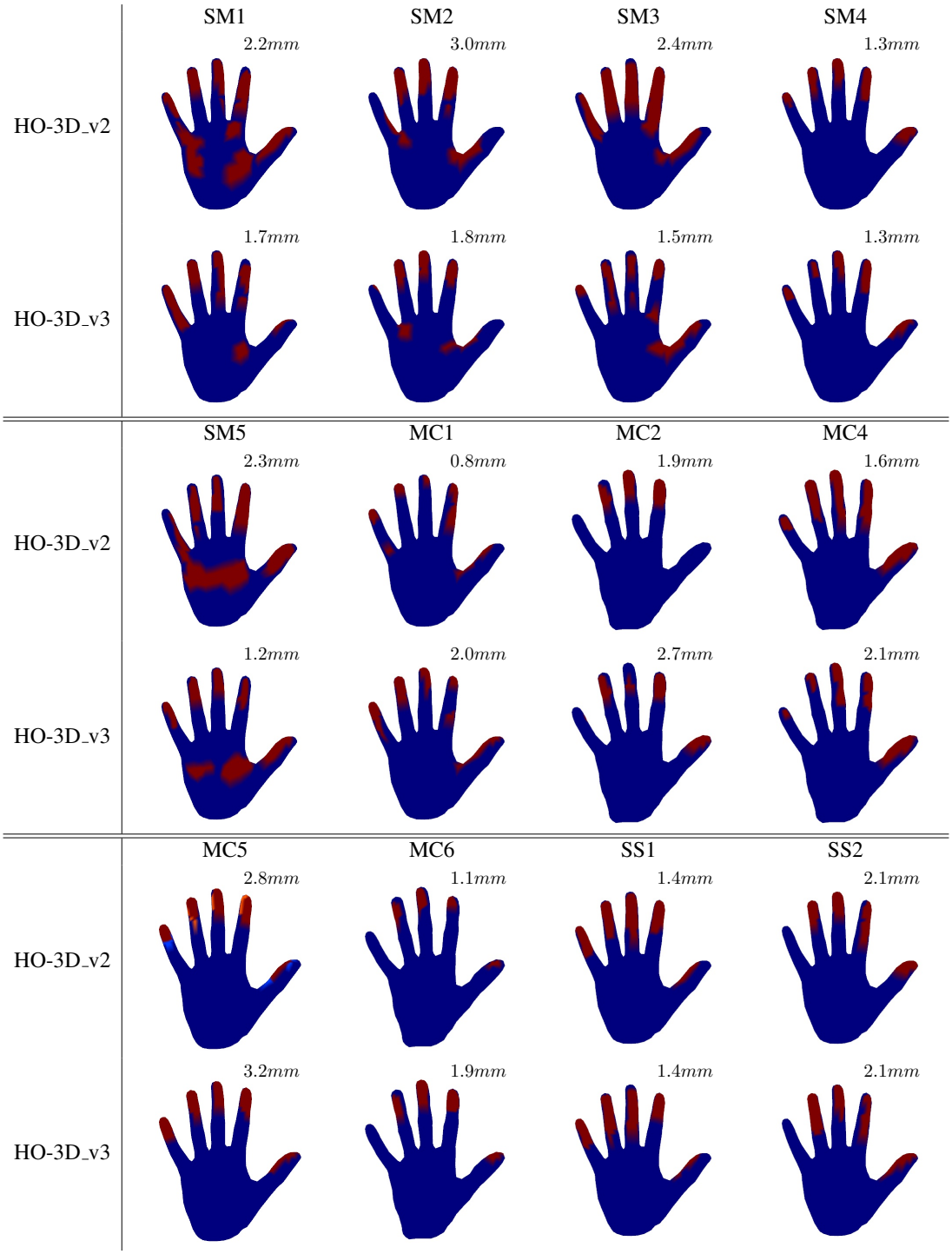


Figure 7: Comparison of contact maps between the 'v2' and 'v3' versions of the HO-3D dataset for the static grasp sequences. The mean penetration for each sequence is provided in the inset. The grasp quality between the two versions are similar.

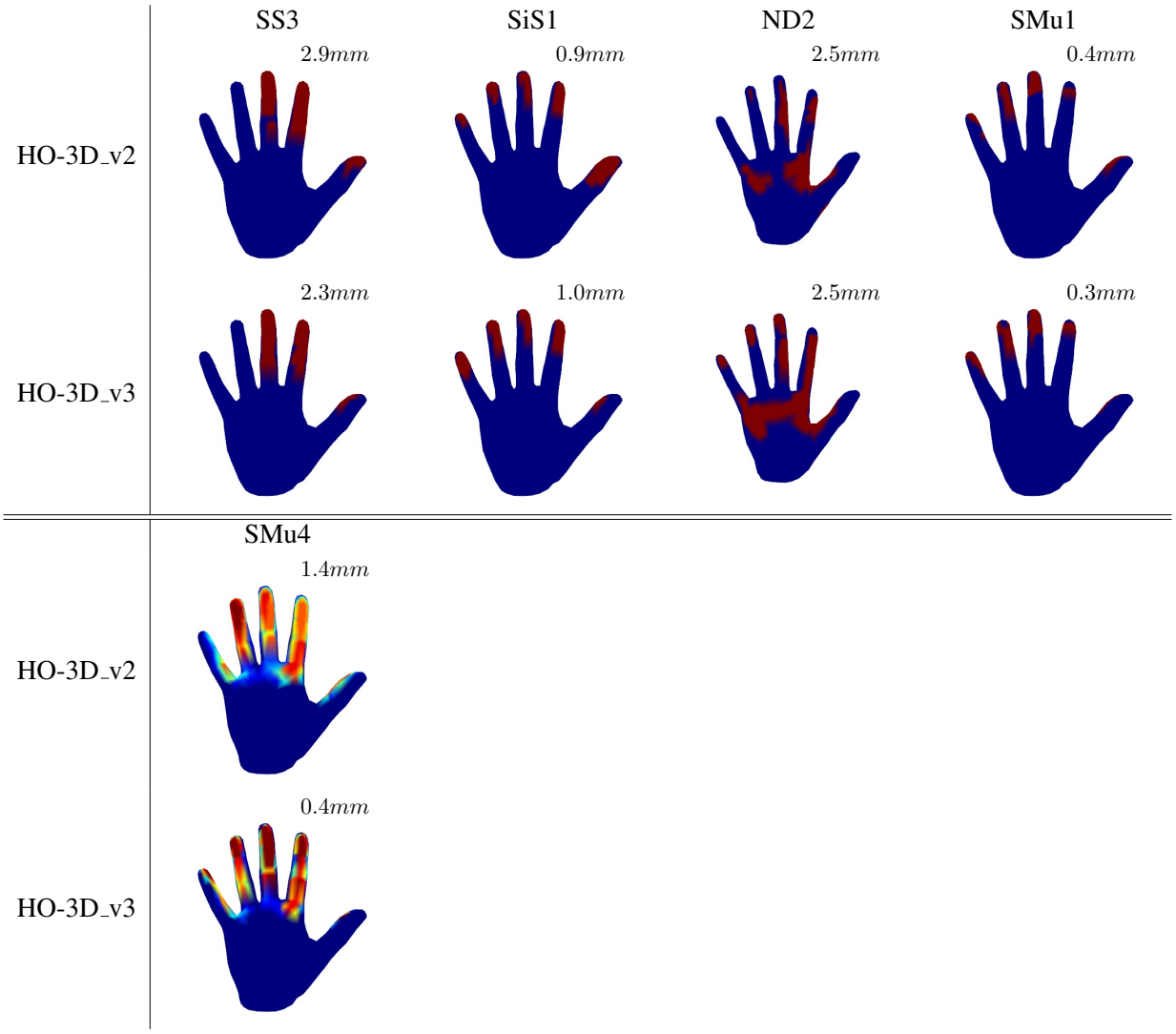


Figure 8: Comparison of contact maps between the 'v2' and 'v3' versions of the HO-3D dataset for the static grasp sequences. The mean penetration for each sequence is provided in the inset. The grasp quality between the two versions are similar.