

Hand-Object Contact Prediction via Motion-Based Pseudo-Labeling and Guided Progressive Label Correction: Supplementary Materials

Takuma Yagi

tyagi@iis.u-tokyo.ac.jp

Md. Tasnimul Hasan

tasnim@iis.u-tokyo.ac.jp

Yoichi Sato

ysato@iis.u-tokyo.ac.jp

The University of Tokyo

Tokyo, Japan

A.1 Details on Pseudo-Label Generation

Preprocessing We extracted frames from videos by either 30 or 25 fps, half of the original frame rate. We processed all the frames in the resolution of 854×480 .

Hyperparameters In the implementation, we used different σ and motion thresholds for contact detection and no-contact detection. We summarize the hyperparameters used in Table 1. d_h and d_o denote the average motion direction of the hand region and object region. We tuned the hyperparameter using the validation set.

For pseudo-label extension, we tracked at most 100 points for each hand and object to track the distance between hands and objects.

Additional examples Figure 1 shows additional results on pseudo-label generation. As seen in the figures, our procedure assigns reliable pseudo-labels in various types of interactions. However, few tracking errors are included (e.g., rightmost frame in second example) and the label assignment is not perfect, suggesting the needs of correction.

A.2 Model Details

Network architecture Figure 2 shows the architecture of the proposed contact prediction model. The input frames are passed one by one and temporal dependencies will be captured at the bidirectional LSTM layers.

Training During training, if the length of the hand-object track is long, we randomly cropped the track at a maximum length of 105 to fit the GPU memory.

	σ	hand	object	background	motion direction
Contact	2.0	$h_r \geq 0.7$	$o_r < 0.05$	$b_r < 0.2$	$\text{sim}(d_h, d_o) > 0.5$
No-contact	1.0	$h_r \geq 0.7$	$o_r \geq 0.2$	$b_r < 0.2$	$\text{sim}(d_h, d_o) < 0.0$

Table 1: Hyperparameters used for pseudo-label generation.

Amount of pseudo-labels	Frame Acc.	Boundary Score	Peripheral Acc.	Edit Score	Correct Ratio
0% (supervised-train)	0.770	0.563	0.649	0.718	0.394
1%	0.784	0.595	0.728	0.729	0.397
5%	0.803	0.620	0.737	0.747	0.427
25%	0.818	0.651	0.725	0.772	0.467
100% (proposed)	0.836	0.681	0.730	0.793	0.519

Table 2: Ablations on noisy dataset size.

Baseline models In **ContactHands** and **Shan-***, we used the pre-trained model provided by the authors. We used the combined model in the former and the model trained on the 100DOH dataset and egocentric datasets for the latter. In those baseline models, we link between the predicted hand instance mask and ground truth hand instance mask if IoU is above 0.5. In **Shan-Bbox**, we predicted as a contact if IoU between a predicted object bounding box and a ground truth bounding box is larger than 0.5. **Shan-Full** combines **Shan-Bbox** and **Shan-Contact** based on the following rules: (i) If IoU between input hand instance mask and input object bounding box is zero, predicts as a no-contact; (ii) If **Shan-Bbox** predicts as a no-contact, follow its prediction; (iii) Otherwise, use predictions produced by **Shan-Contact**. We observed improved performance by combining predictions based on object-in-contact detection and contact state prediction.

A.3 Dataset Details

Pseudo-labels We used 96,000 tracks (9 million frames) with bounding boxes and pseudo-labels for training. Pseudo-labels were assigned for 37.3% of the total frames.

Trusted labels We annotated 67,064 frames of 1,200 tracks. We did not annotate the instance masks of the hands since the segmentation network described in Section 3.1 produced reliable results. The average length of the track was 56 frames. To evaluate whether the model can distinguish touched and untouched objects, we included tracks stably in contact and untouched tracks. The number of tracks that were in constant contact was 284, the number of tracks with mixed contact states was 670, and the number of tracks that were not in contact was 246.

A.4 Additional Experimental Results

Effect of pseudo-label set size. Table 2 shows ablation results on changing the noisy dataset size. We sampled 1%, 5%, and 25% of the full noisy dataset and trained by the proposed gPLC algorithm. The result supports the fact that using large-scale data with pseudo-labels helps generalization.

Additional qualitative examples Figure 3 shows additional examples on the contact prediction result. Our model focused on motion rather than spatial overlap to infer the contact

states. The performance was boosted at novel objects thanks to large-scale pseudo-label training.

We also show additional failure cases in Figure 4. We observed failures due to unfamiliar grasps (top), no movement (middle), subtle hand movement (bottom). Since we focused on the holistic foreground motion of hands and objects, it was difficult to predict contacts in fine-grained manipulation.

Effect of input modality Figure 5 shows the prediction examples on different input modalities. In general, the model trained by RGB input solely tended to make uncertain predictions near boundaries (see Figure 5 top). Also shown in the lower boundary score, this result indicates distinguishing a contact state is difficult from a single image. The model trained by flow input solely generally behaves similar to the proposed model. However, the difference appears when there is no motion in the scene. If there is no or subtle motion in the scene, the flow model has no clue except temporal context to predict contacts. In such cases, RGB images will be the only clue (see Figure 5 middle). Motion information contributed to accurate prediction in most cases but sometimes failed when complex motion patterns are observed (see Figure 5 bottom).

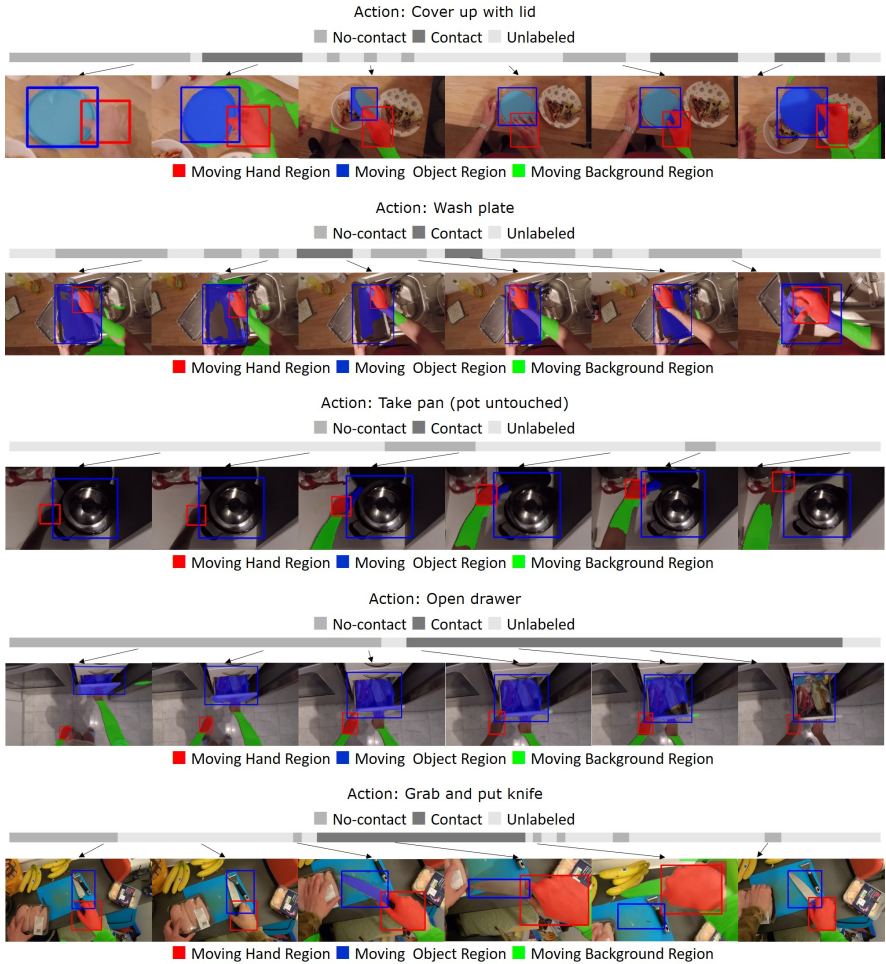


Figure 1: **Additional examples of generated pseudo-labels:** (Top) Gray and dark gray bar indicates no-contact/contact labels otherwise no labels assigned. (Bottom) Representative frames. Red, blue, and green regions denote moving hand, object, and background regions, respectively. Note that in few tracking errors are included in these tracks (*e.g.*, rightmost frame in second example). Refer to video visualization for detail.

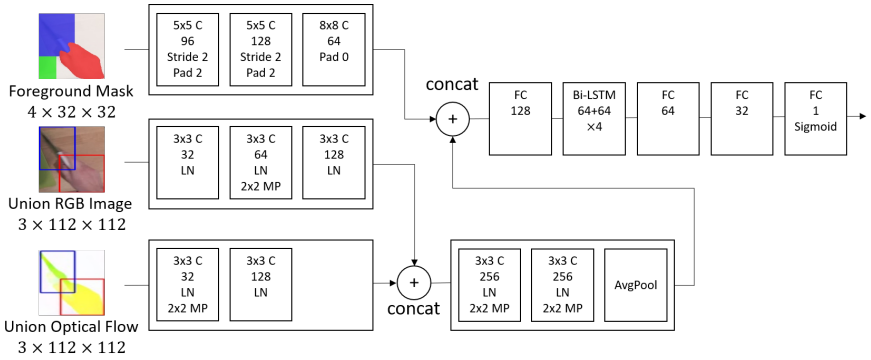


Figure 2: Detailed network architecture of contact prediction model (per frame). C denotes convolutional layer with filter size and number of channels, followed by ReLU layer. MP denotes max pooling with filter size. LN denotes layer normalization layer. FC denotes fully-connected layer with number of units, followed by ReLU layer (except last layer).

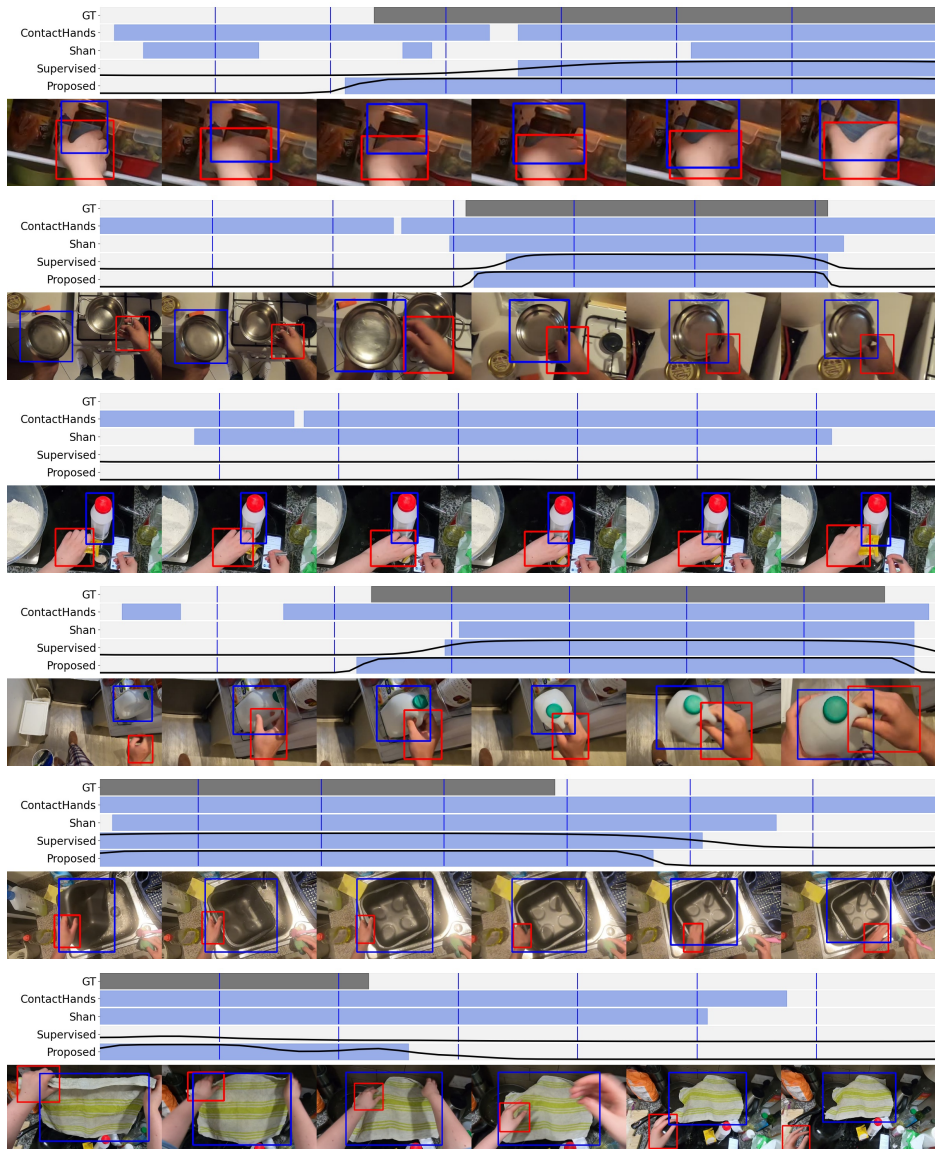


Figure 3: Additional qualitative examples. Our model better predicts correct contact state change point and can avoid false positives in difficult cases of image-level overlap between hand and objects. Refer to video visualization for detail.

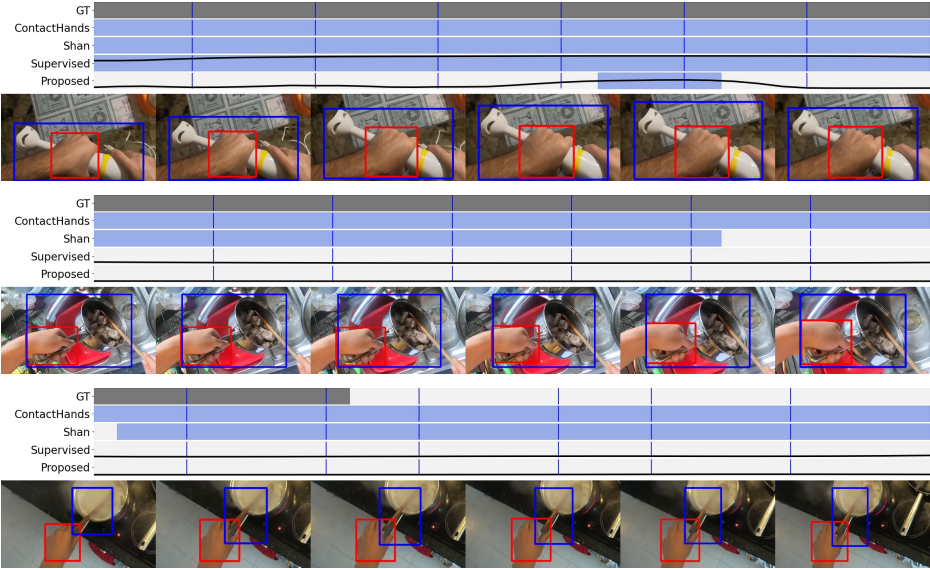


Figure 4: Additional failure examples.

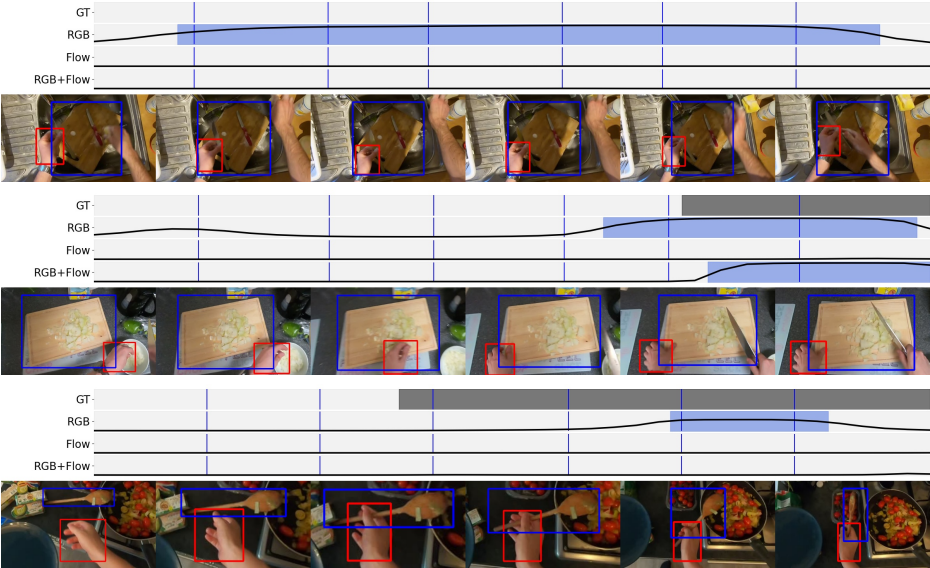


Figure 5: Qualitative results on input modalities.