Learning Visuo-Haptic Skewering Strategies for Robot-Assisted Feeding
Supplementary Material

Datasets, code, and videos can be found on our website.

A Experimental Hardware

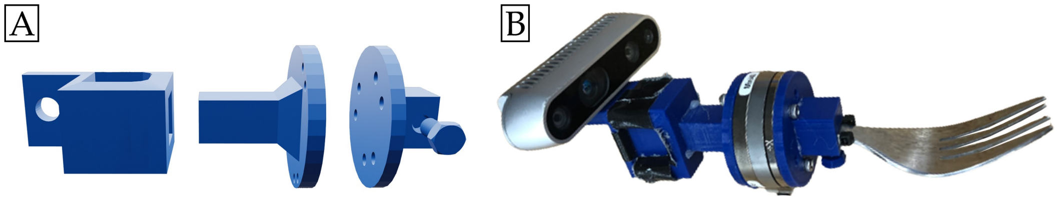

Figure 6: **End-Effector Custom Hardware:** A) CAD designs for custom end-effector attachment with inserts for Franka Panda gripper tips and mounting holes for Intel RealSense D435i, Mini45 ATI F/T sensor, and fork with screw attachment. B) The 3/D printed full mount, assembled using mounting screws and super glue.

We collect all data and run all experiments using the Franka Panda 7DoF robot with the custom 3D-printed end-effector mount shown in Figure 6. While the fork is attached via a screw-in slot, the force of consecutive skewering attempts in the course of clearing a plate can cause the fork position to shift slightly within the slot. To address this, we train ServoNet to estimate the fork-food item offset and continuously servo until the midpoint of the tines is aligned with the predicted item center (Figure 7).

B Additional Qualitative Results

In this section, we provide additional qualitative results of the full bite acquisition pipeline: RetinaNet bounding box detection, SkeweringPoseNet for item pose estimation, ServoNet for pre-probing, and HapticVisualNet for primitive planning over consecutive actions.

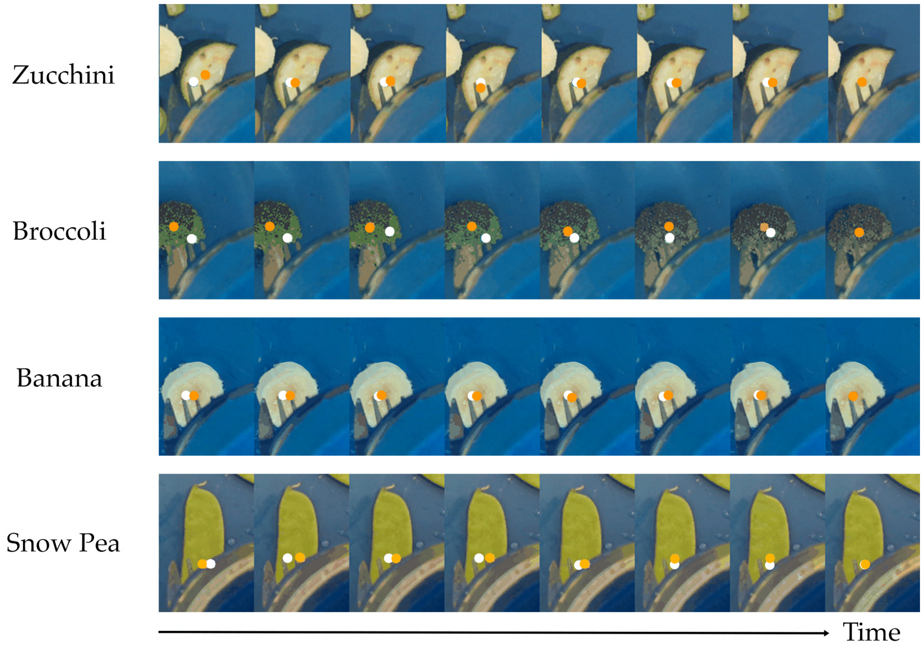

Figure 7: **ServoNet**: From left to right, we plot the last 8 frames of visual servoing using ServoNet until the fork tines midpoint (white) is centered with the predicted item location (orange), across 4 items (zucchini, broccoli, banana slice, snow pea).

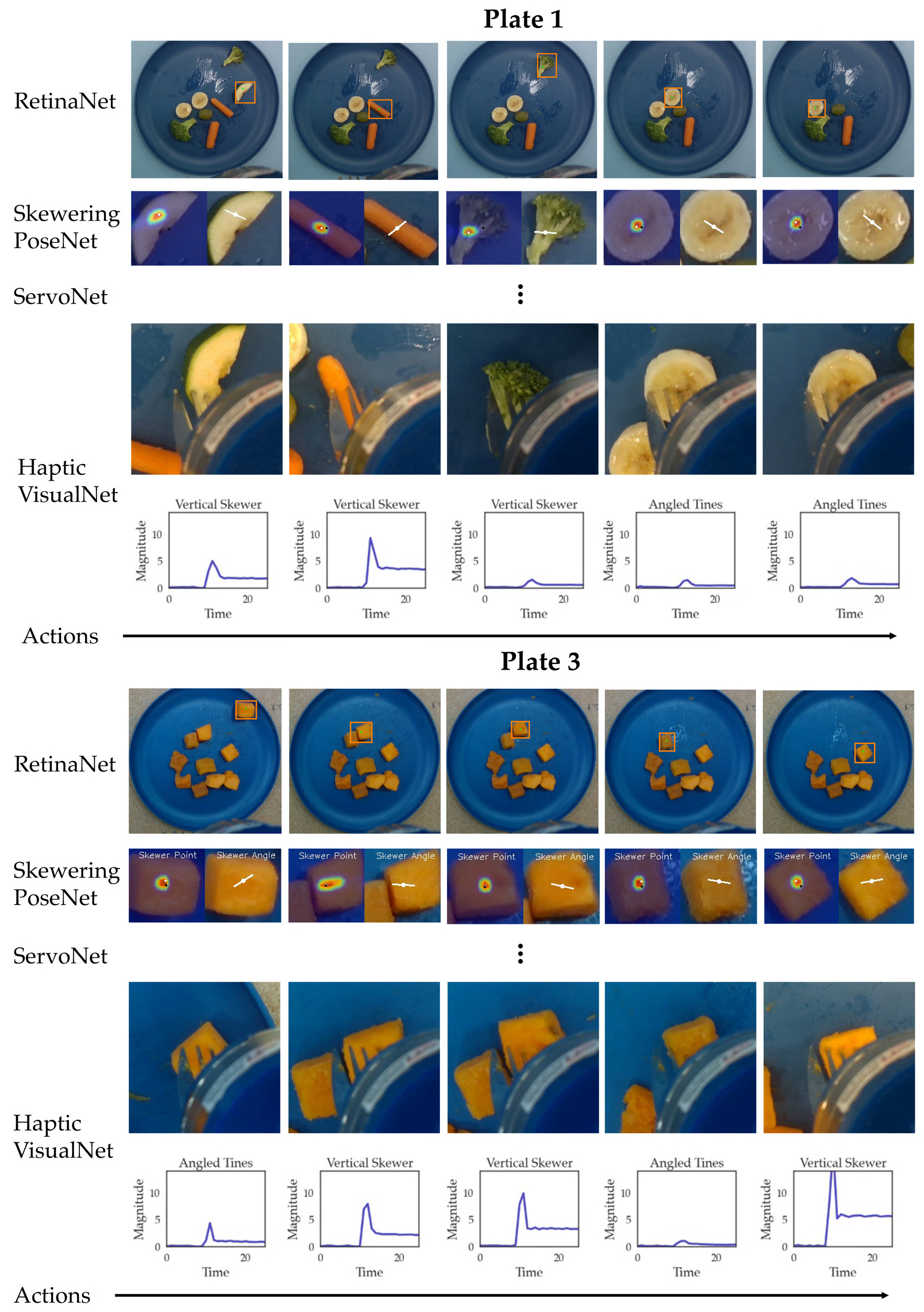

Figure 8: **HapticVis Action Predictions:** We visualize 5 consecutive HapticVis action predictions (left to right) for Plates 1 and Plates 3. The SkeweringPoseNet predictions per column first illustrate the predicted skewering points (which are intended to refine bounding box centers depicted in row 1 to item locations) and secondly the skewering orientations. In Plate 1, HapticVis correctly infers a sequence of actions to pick up zucchini, a raw carrot, a broccoli floret, and two slices of banana. In Plate 3, HapticVis skewers several pieces of butternut squash ranging from raw (1st-3rd, 5th columns) to overcooked (4th column). These intrinstic properties mostly align with the predicted action predictions, except for one mis-predicted `angled skewer` primitive for the first action, which succeeds but results in an unstable skewer. Plate 3 column 2 is a failed action in which ServoNet erroneously guides the fork to the wrong location, likely due to the close proximity of two butternut squash pieces. This causes a drop failure, but the item is re-attempted and successfully skewered in the next action.

C ServoNet Training Details

We employ ServoNet, a network which continuously estimates the fork-food offset from images, and precisely guides the fork to a desired item via visual servoing in order to probe. We implement ServoNet with a fully-convolutional ResNet-18 backbone. ServoNet takes as input a 200 × 200 × 3 image from the end-effector mounted RealSense camera, and outputs a 200 × 200 × 2 heatmap. The two channels represent a 2-d Gaussian heatmap centered around the predicted fork pixel and nearest food pixel, respectively. We take the predicted fork and food pixels to be the argmaxes of these heatmaps.

To train ServoNet, we annotate 200 images with the corresponding fork and nearest food item pixels and augment this dataset to 3,500 examples (Figure 9). We train the network using binary cross-entropy loss over the predicted and ground truth heatmaps. Figure 10 shows some visualizations of the predicted fork-food heatmaps on unseen images.

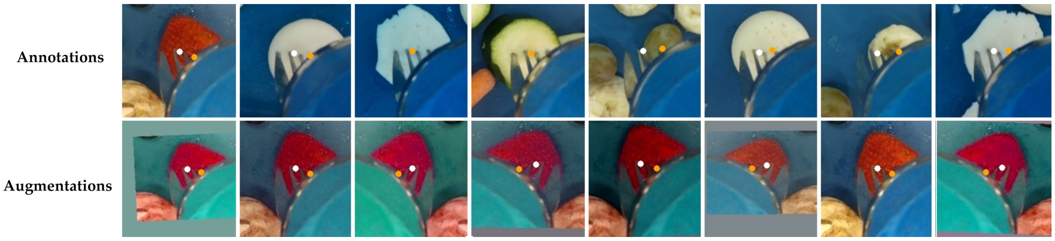

Figure 9: **ServoNet Dataset Generation:** Top row: we annotate 200 images with the fork tines center (white) and nearest food item center (orange). Bottom row: we augment this dataset with various colorspace and affine transformations to yield a dataset of 3,500 examples (shown for the first strawberry example).

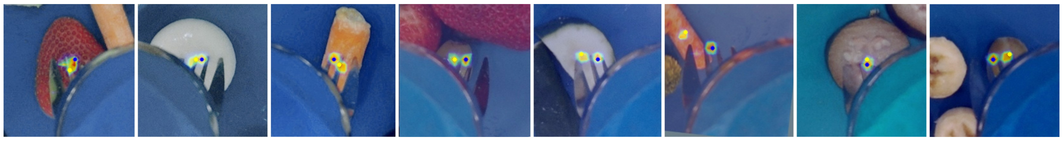

Figure 10: **ServoNet Predictions:** For 8 unseen images, we visualize the ServoNet Gaussian heatmap predictions. The estimated fork pixels appear in blue, and the estimated food item pixels are shown in yellow.

D Additional Ablations

D.1 Sample Efficiency

The original dataset collected to train HapticVisualNet consists of 300 examples of paired post-contact images, force readings, and manually assigned primitive labels, augmented 8x artificially (Section 3.5). To understand the sample efficiency of HapticVisualNet, we ablate for the accuracies of the network when training on varied amounts of data.

Training Dataset Size (% of orig.)	Overall Acc.	Vertical Skewer Acc.	Angled Skewer Acc.
100% (OURS)	**94.1**%	**95.2**%	93.1%
75%	91.8%	88.3%	95.7%
50%	92.1%	87.6%	**97.8**%
25%	89.4%	83.4%	96.9%
10%	84.2%	82.7%	86.1%

Table 2: **HapticVisualNet Accuracy and Sample Efficiency:** Training with all 300 of the data points, augmented, yields the highest overall accuracy. We note a general trend towards lower accuracies as the dataset size decreases, taking into account that with smaller dataset sizes comes lower state coverage and higher variance in the accuracies, resulting in the 50% network achieving high angled skewer accuracy but low vertical skewer accuracy.

D.2 Interpretability of HapticVisualNet

Given a paired food image observation and haptic readings, HapticVisualNet separately encodes visual features and haptic features. The concatenated features yield a 640-d multimodal embedding.

This embedding serves as input to a final linear layer followed by a softmax activation which yields predicted primitive likelihoods. To improve interpretability of HapticVisualNet's learned food item representation, we visualize the multimodal embeddings for 26 items in the training dataset using 2-d t-SNE projections below.

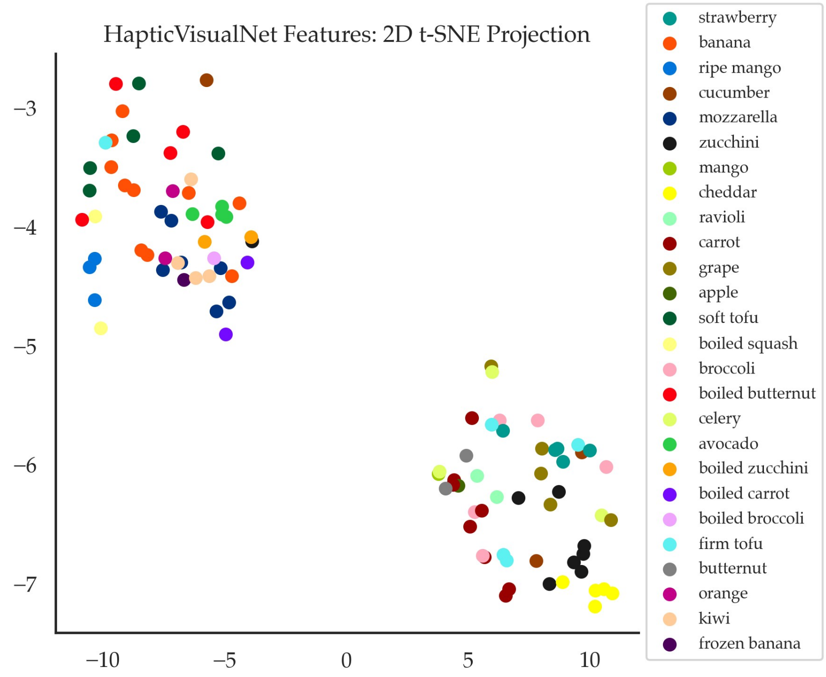

Figure 11: **2D t-SNE Projection of HapticVisualNet Embeddings**: We find that instances of the same item (i.e. ripe mango) and items with similar textural properties (i.e. boiled carrot and boiled zucchini, avocado and banana) tend to cluster. Additionally, HapticVisualNet learns to separate items that favor `vertical skewer` (bottom right) from those that favor `angled skewer` (upper left) by pushing them apart in latent space (i.e. mozzarella vs. cheddar, soft tofu vs. firm tofu).

E Additional Physical Experiments

In these set of experiments, we stress-test HapticVisualNet's capabilities to generalize to challenging textural and visual food properties. We assess HapticVisualNet on the tasks of clearing a plate of frozen fruits and a plate with sauteed vegetables and tofu in sauce.

Items	Skewering Success Rate	Slip/Miss	BBox FP	3+ Tries
Frozen mango, pineapple, strawberry	20/25 [video]	5	0	1
Sautéed veggies and tofu	20/23 [video]	3	3	0
Sautéed veggies, tofu, soy sauce	25/34 [video]	9	6	1

Table 3: **HapticVisualNet Stress-Tests:** HapticVisualNet successfully skewers frozen fruit in 20/25 and stir-fried tofu with vegetables in 45/57 total attempts. Videos of both experiments are available on the website. The failures observed in the fruit case are slips during probing due to the highly rigid texture of frozen fruits (and in the case of one strawberry, consecutively slipping three times on the same item). In this experiment, we also note that HapticVisualNet predominantly infers vertical skewering for these items as expected, but occasionally predicts angled tines. This is possibly due to the effects of thawing and softening over time. In the sautéed vegetables experiment, HapticVisualNet has more near misses and slips due to the degree of clutter and oiliness of the surface. As these items are very out of distribution, this also incurs more bounding box failures leading to early termination. Still, HapticVisualNet demonstrates generalization to charred/oiled/sauce-coated/seasoned items of varying levels of doneness.