Supplementary Materials

Behavior	Total	Average Per Video
Lift	1175	1.01
Hand-open	2227	1.91
Grab	2096	1.79
Supinate	1392	1.19
At-mouth	921	0.79
Chew	664	0.57
Background	830939	71081

Table 1. Number of labelled frames with the Mouse Reach Dataset.

Mouse	Total Videos
M134	217
M147	97
M173	492
M174	359

Table 2. The Mouse Reach Dataset contains a total of 1169 videos of mice performing the reaching task.

Table 1 and Table 2 provide more details on our dataset. The behaviors: Hand, Grab and Supinate, occur more often because the mouse will fail to grab the food pellet and try to grab food pellet again. The number of chew frames are low because the mouse will also fail to eat the food pellet. Figure 2 and Figure 3 show sample frames of each of the behaviors.

Fig. 6 shows a diagram of our base model. The base model is a two layer bi-directional LSTM with 256 hidden units. The inputs to the LSTM pass through a fully connected layer, ReLU, and Batch Normalization. The outputs are transformed by a fully connected layer with a sigmoid activation layer.

Fig. 4 and Fig. 5 show more examples of the distribution of predictions for each behavior. For most of the behaviors, the number of false positives within $\tau = 10$ frames is greatest for the MSE loss.

Fig. 7 shows a larger screenshot of our visualization tool. The web-based viewer synchronizes the network output score and video frame. The viewer has two main components. A line graph, where the x-axis is video frames and the y-axis is the network output score. The other component is a video viewer, where the frame being shown is the currently selected frame. A frame can be selected by either playing the movie or mousing over the line graph. Being able to slowly, or quickly, mouse over consecutive video frames around curious network outputs helped find software bugs and explore network architectures.

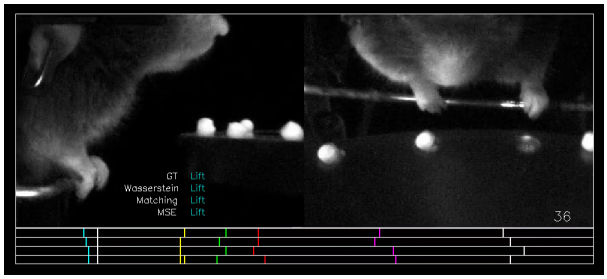

Figure 1. Example frame from a video showing the classification results. See text for details

We also provide four videos showing examples of the reach task with the labels predicted by our trained models. Each video shows the front and side views of the mouse with four rows of bars beneath the mouse frames. Figure 1 shows an example frame snapshot. The first row represents the ground truth location of the behaviors, and the following three rows represent the Wasserstein, Matching, and MSE model predictions. Each row has colored vertical bars for each of the behaviors. "Lift" is cyan, "Hand" yellow, "Grab" green, "Supinate" red, "At-mouth" magenta, and "Chew" white. A vertical bar representing the current frame will move across the four rows. Additionally, within the video frame of the mouse reaching task, we add text labels of the predictions. As the video playback frame approaches the frame location of a predicted label, the label name will fade in, and fade out as the playback passes by. The example videos of action start detection are available at `http://research.janelia.org/bransonlab/MouseReachData/`. M134_20150325_v020.mp4, M174_20150417_v031.mp4, and M174_20150427_v004.mp4 show the mouse successfully grab and eat the food pellet. In videos M134_20150325_v020.mp4 and M174_20150427_v004.mp4, all three networks properly predict each behavior, however the MSE loss produces extra false positives. In M174_20150417_v031.mp4, all three networks struggle to properly detect each behavior. Both the Matching and MSE losses produce extra false positives, while the Wasserstein loss fails to detect the

”Chew” behavior. M134_20150504_v018.mp4 shows an example of the mouse failing to grab the food pellet on its first try. The Wasserstein loss properly detects each behavior, while the MSE loss produces a large number of false positives.

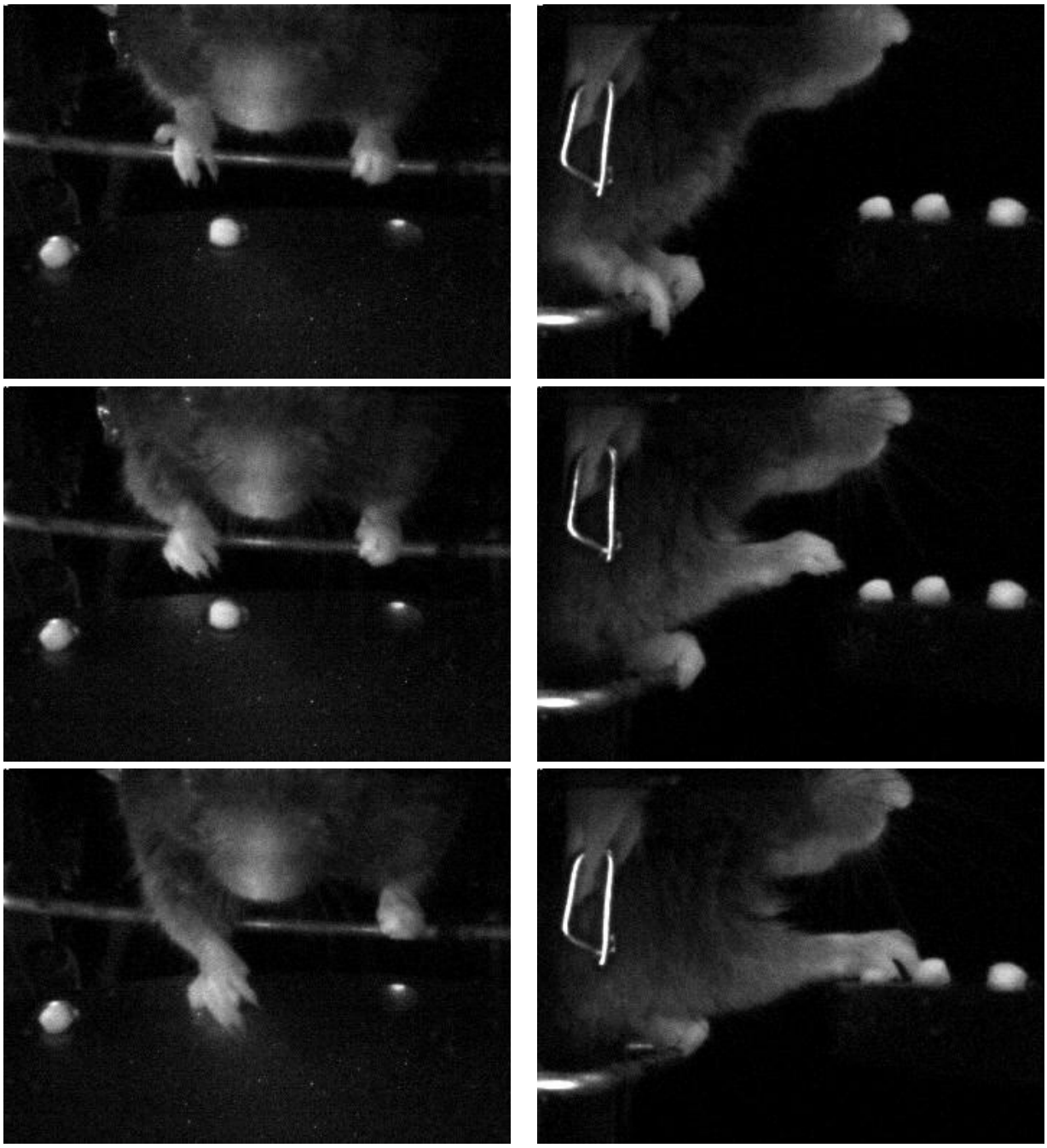

Figure 2. Example frames of behaviors the Mouse Reach Dataset. The first row is the Lift behavior. Here the mouse paw is beginning to move off of the perch. The next row is the Hand-open behavior. Here is the mouse beginning to open his paw to grab a pellet. The third row is the Grab behavior. The mouse beginning to close his paw around a food pellet.

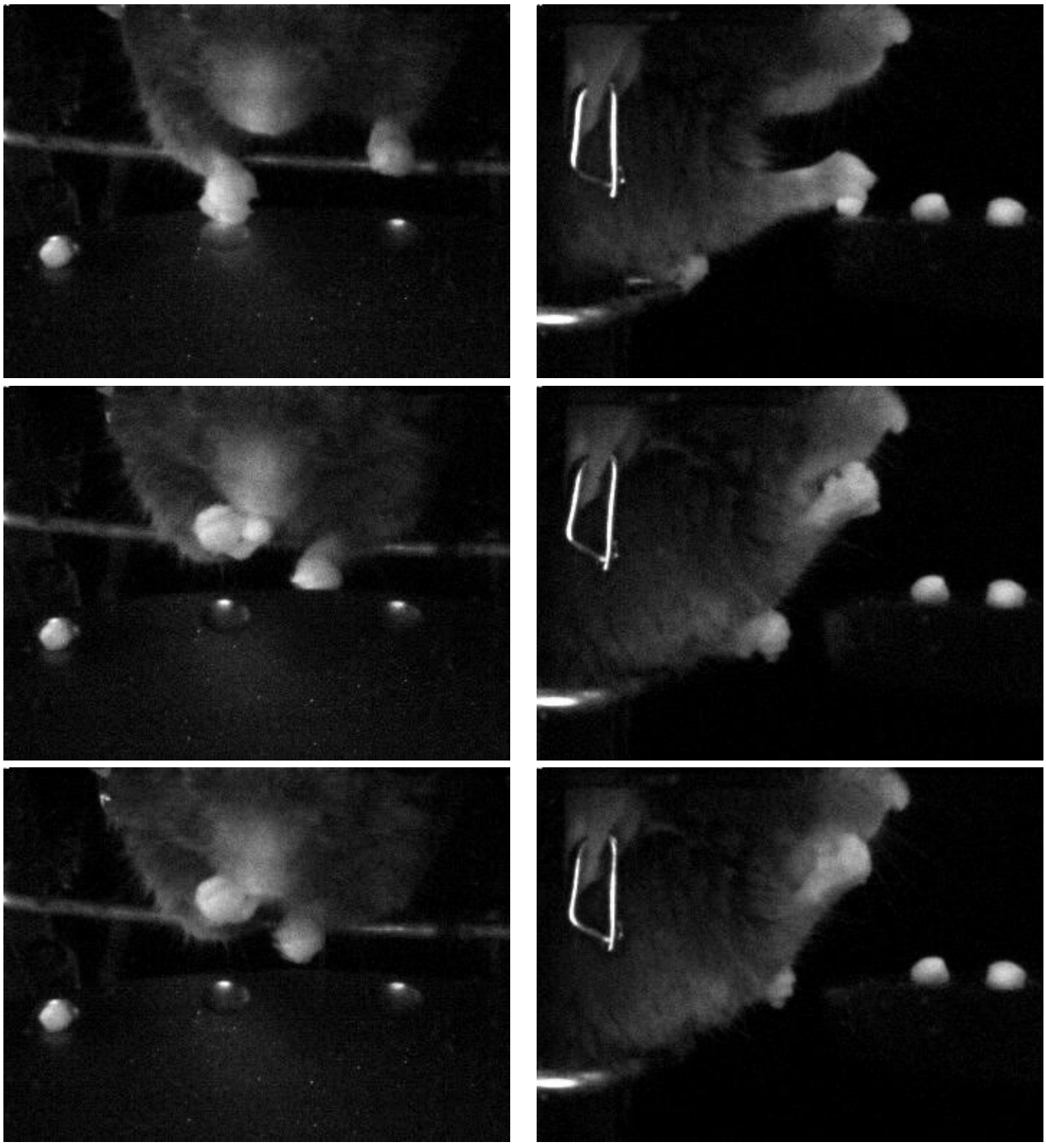

Figure 3. The Supinate behavior is shown in the first row. The mouse is beginning to turn its paw towards its mouth. The second row shows the At-mouth behavior. The mouth behavior occurs when the food pellet is starting to be placed into the mouth. The last row shows the Chew behavior, where the food pellet in the mouth and the mouse is starting to eat the pellet.

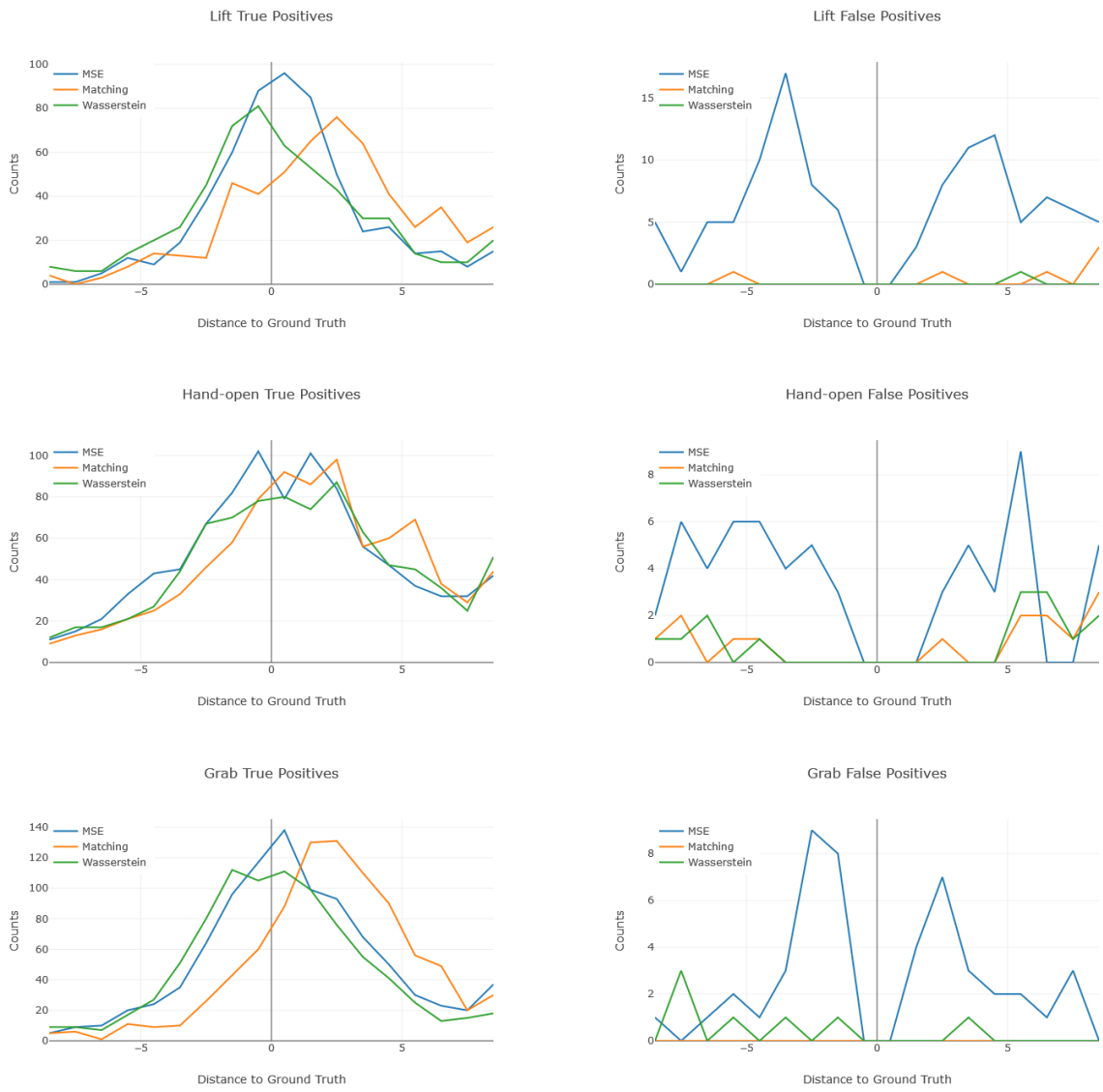

Figure 4. Left column shows the distribution of true positives and the right side the false positives. For these behaviors the network is able to localize the start frame accurately.

Figure 5. Left column shows the distribution of true positives and the right side the false positives. For these behaviors, the network struggles to predict the start frame accurately.

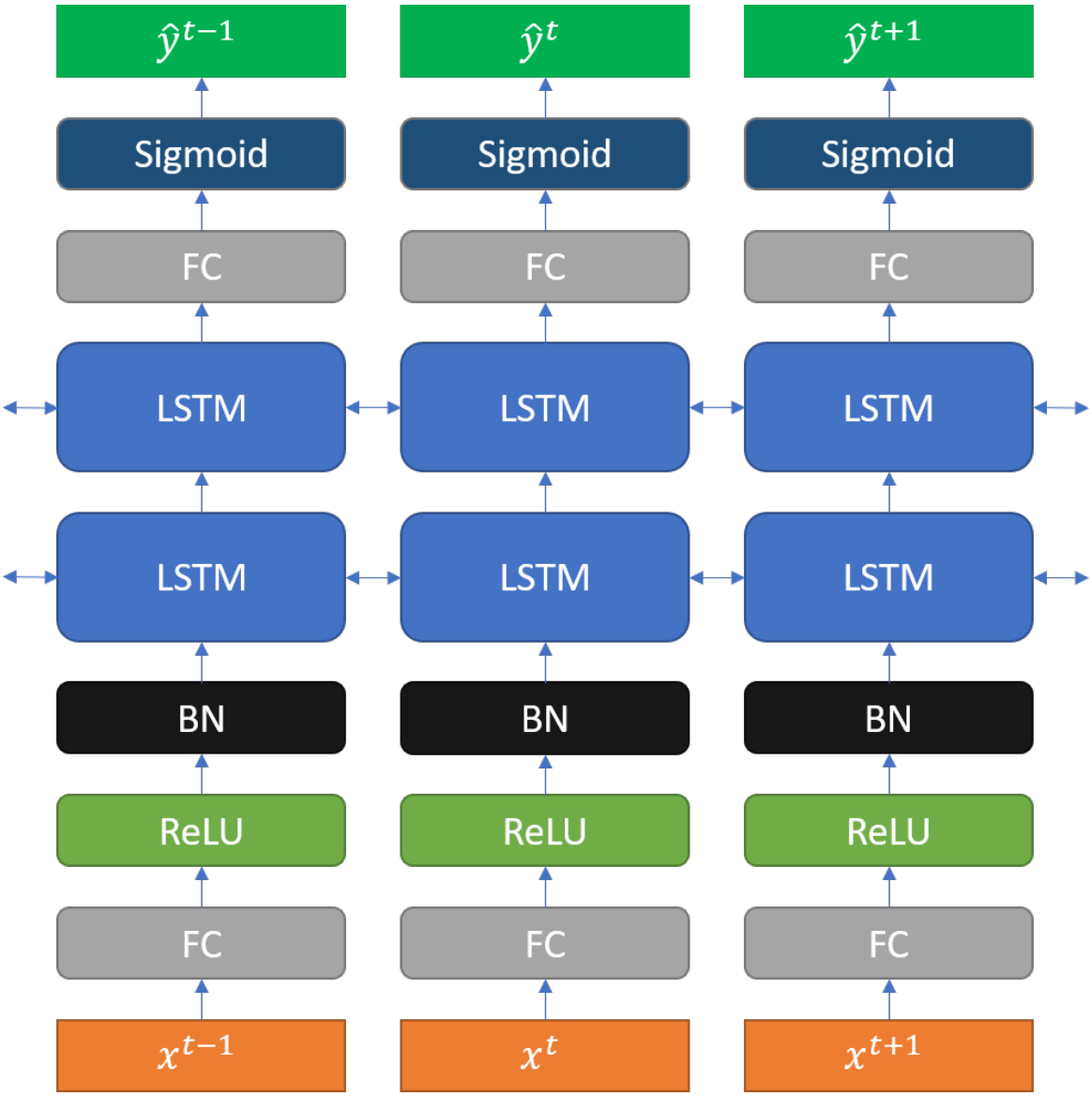

Figure 6. Our complete model consists of a fully connected layer, ReLU, Batch Normalization, two Bi-directional LSTM layers, a fully connected layer then a sigmoid activation layer. The LSTMs each have 256 hidden units.

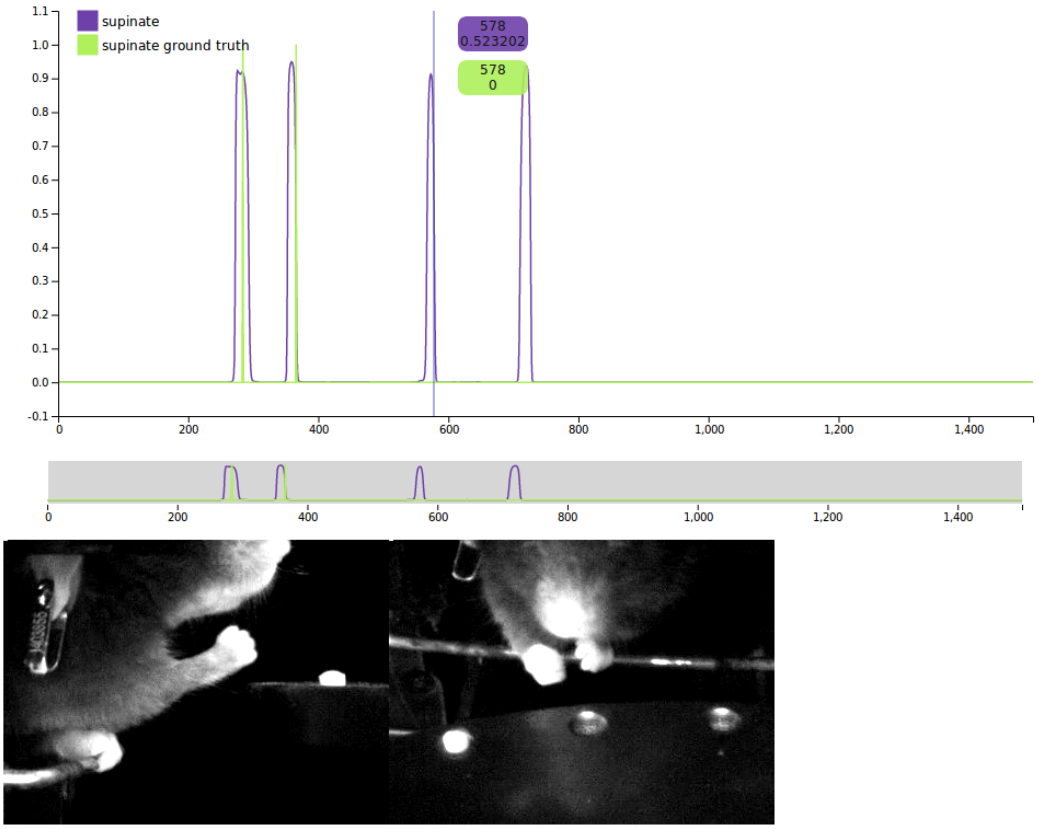

Figure 7. An example screen shot of our web based network output viewer for videos. The green line is ground truth and purple is our network's predictions. Here we can mouse over the frames that caused the false positive predictions. The veritical blue line near 600 frames denotes the current visible frame in the video. The purple and green rectangles shows the frame number and scores of that frame. In this case, frame 578 is being viewed and the ground truth supinate score is 0 and the network prediction of the behavior is 0.52. We can see the side of the mouse paw, which is something visible in the mouse supinate behavior, but the paw is quite far from the food pellet.