# SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again Supplementary material

## Errata

After submission, we realized that one value in the detections scores graph for the multi-object dataset (Figure 7, left) was still from an earlier, erroneous run. Specifically, the correct recall for threshold 0 is 0.923 instead of 0.982. We apologize for this mistake but believe that it does not diminish the overall novelty and results of our work.

## 1. Object-wise detection scores

We present the detection score graphs for each object of the first two datasets in Figures 1 and 2 from which we determined the best object-wise threshold. For reproducibility, we list them in Tables 1 and 2.

| Camera | Coffee | Joystick | Juice | Milk | Shampoo |
|---|---|---|---|---|---|
| 0.55 | 0.35 | 0.5 | 0.25 | 0.3 | 0.45 |

Table 1: Object-wise thresholds for the Tejani dataset.

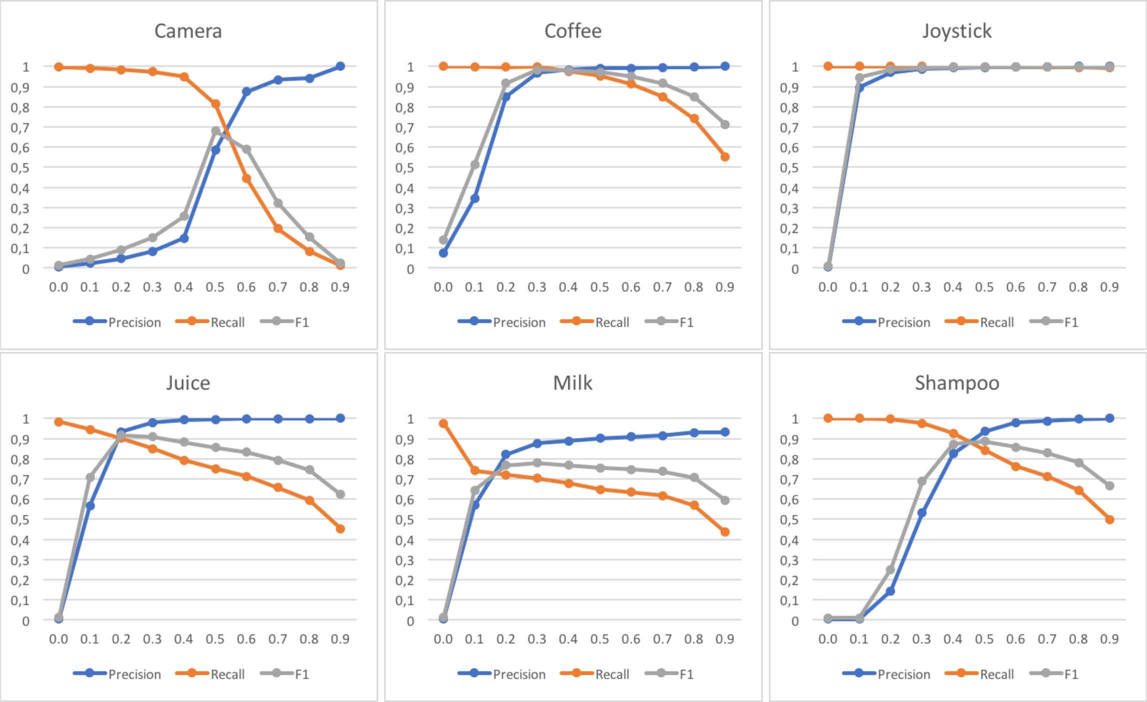
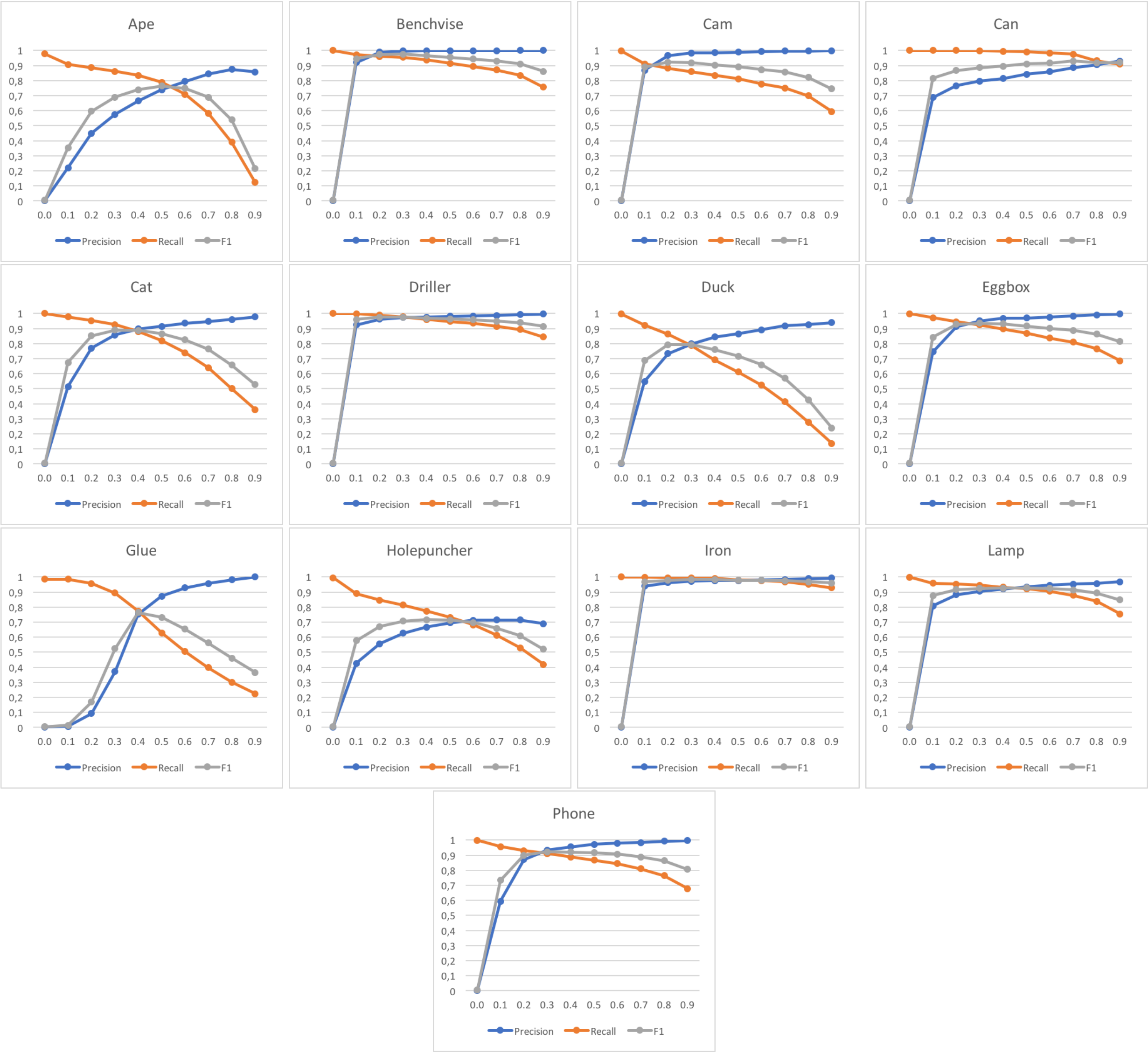


Figure 1: Plotting the detection scores for each object on the Tejani dataset for a varying threshold.

| ape | bvise | cam | can | cat | driller | duck | box | glue | holep | iron | lamp | phone |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.5 | 0.15 | 0.2 | 0.75 | 0.35 | 0.25 | 0.25 | 0.25 | 0.4 | 0.4 | 0.3 | 0.55 | 0.35 |

Table 2: Object-wise thresholds for the LineMOD dataset.


Ape

Benchvise

Cam

Can

Cat

Driller

Duck

Eggbox

Glue

Holepuncher

Iron

Lamp

Phone

Figure 2: Plotting the detection scores for each object on the LineMOD dataset for a varying threshold.

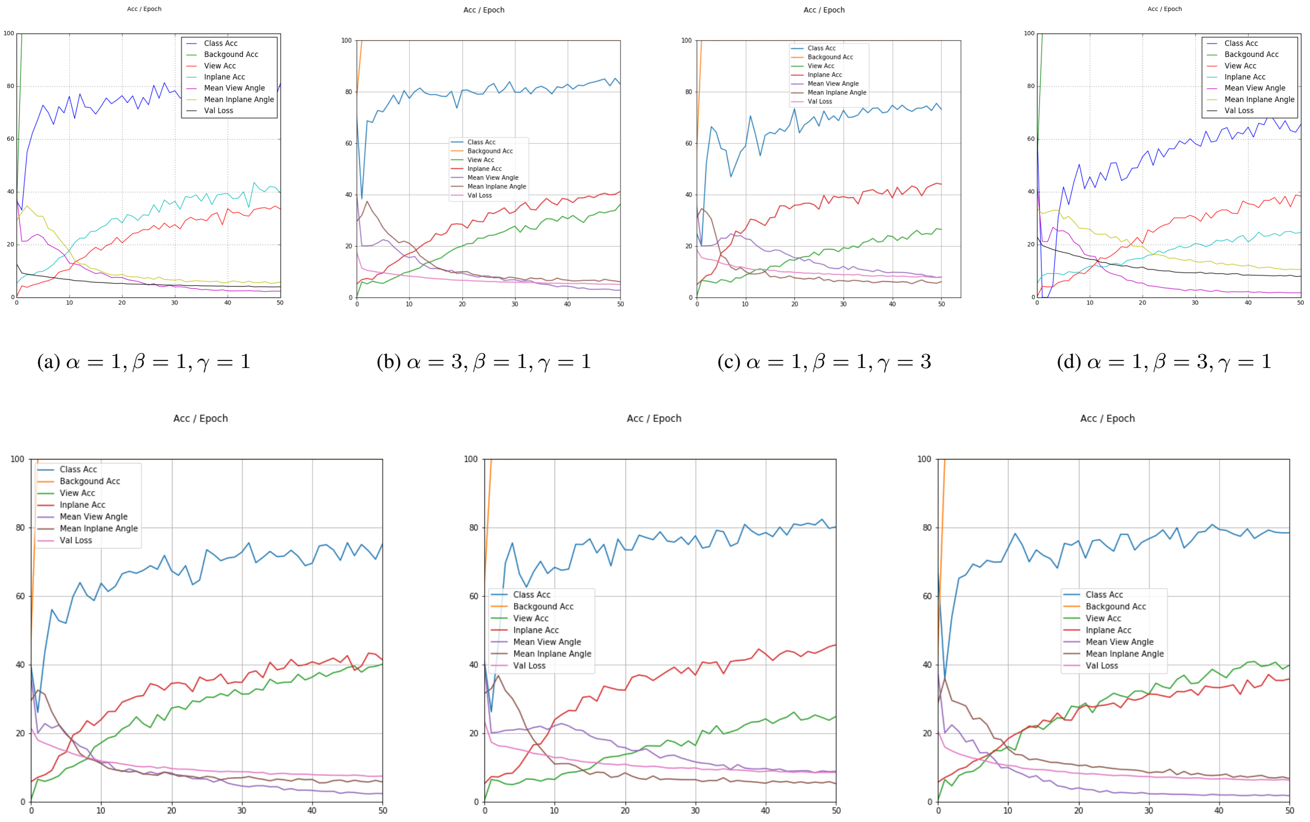
## 2. Detailed pose errors for the LineMOD dataset

| | ape | bvise | cam | can | cat | driller | duck | box | glue | holep | iron | lamp | phone |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IoU-2D | 0.99 | 1.00 | 0.99 | 1.0 | 0.99 | 0.99 | 0.98 | 0.99 | 0.98 | 0.99 | 0.99 | 0.99 | 1.00 |
| IoU-3D | 0.96 | 0.98 | 0.98 | 0.99 | 0.95 | 0.95 | 0.95 | 0.98 | 0.89 | 0.97 | 0.97 | 0.98 | 0.93 |
| VSS-2D | 0.73 | 0.67 | 0.73 | 0.75 | 0.67 | 0.66 | 0.71 | 0.78 | 0.72 | 0.70 | 0.74 | 0.66 | 0.72 |
| VSS-3D | 0.84 | 0.88 | 0.90 | 0.86 | 0.81 | 0.84 | 0.83 | 0.88 | 0.75 | 0.77 | 0.85 | 0.84 | 0.81 |
| ADD-2D | 0.65 | 0.80 | 0.78 | 0.86 | 0.70 | 0.73 | 0.66 | 1.00 | 1.00 | 0.49 | 0.78 | 0.73 | 0.79 |
| ADD-3D | 0.85 | 0.94 | 0.94 | 0.94 | 0.86 | 0.85 | 0.82 | 1.00 | 1.00 | 0.73 | 0.95 | 0.87 | 0.87 |

Table 3: Object-wise pose errors for the LineMOD dataset.

## 3. Error development for different loss term weights

We plot the average error on a synthetic validation set. While the accuracies for class, viewpoint and in-plane rotations increase, the networks converge at different levels. We also plot the more important mean angular deviation for viewpoint and in-plane rotation since this is usually the expected error of the pooled hypotheses before refinement.



(a) $\alpha = 1, \beta = 1, \gamma = 1$

(b) $\alpha = 3, \beta = 1, \gamma = 1$

(c) $\alpha = 1, \beta = 1, \gamma = 3$

(d) $\alpha = 1, \beta = 3, \gamma = 1$

(e) $\alpha = 1, \beta = 2, \gamma = 2$

(f) $\alpha = 3, \beta = 1, \gamma = 3$

(g) $\alpha = 2, \beta = 2, \gamma = 1$

Figure 3: Development of training error on a synthetic validation set.

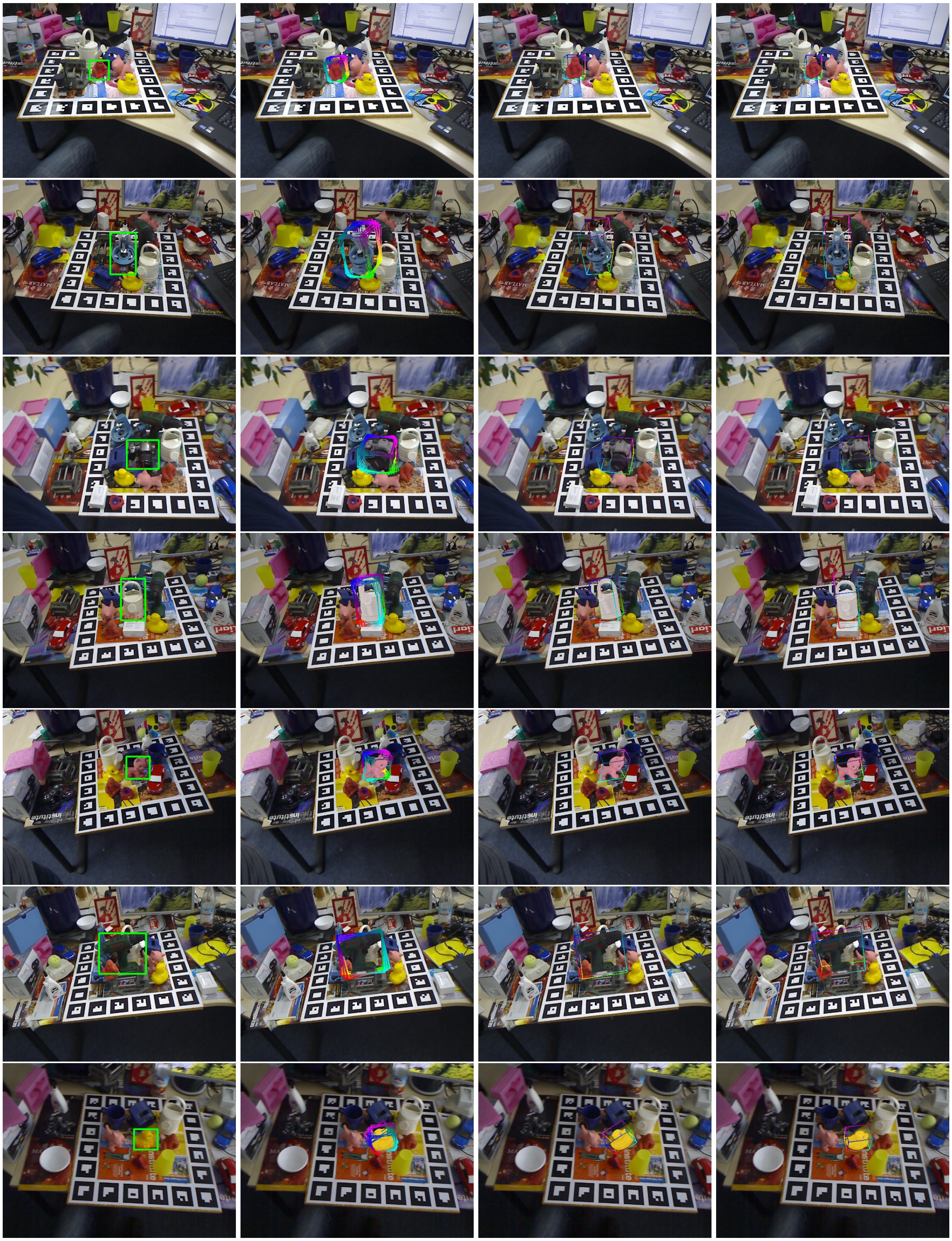


Figure 4: Qualitative results on the LineMOD dataset. From left to right: 2D prediction, hypothesis pool, result after 2D refinement, result after 3D refinement.

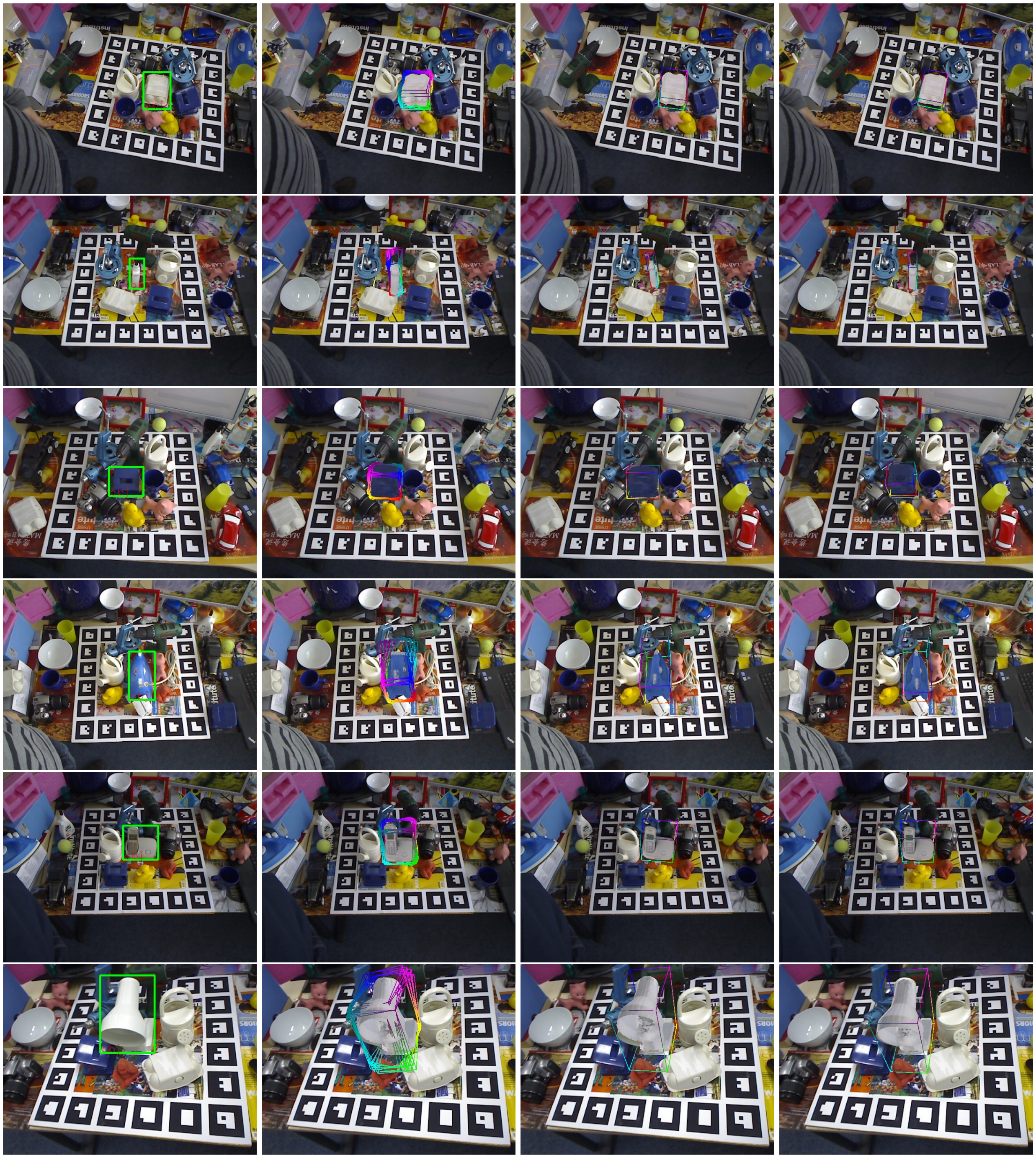


Figure 5: Qualitative results on the LineMOD dataset. From left to right: 2D prediction, hypothesis pool, result after 2D refinement, result after 3D refinement.

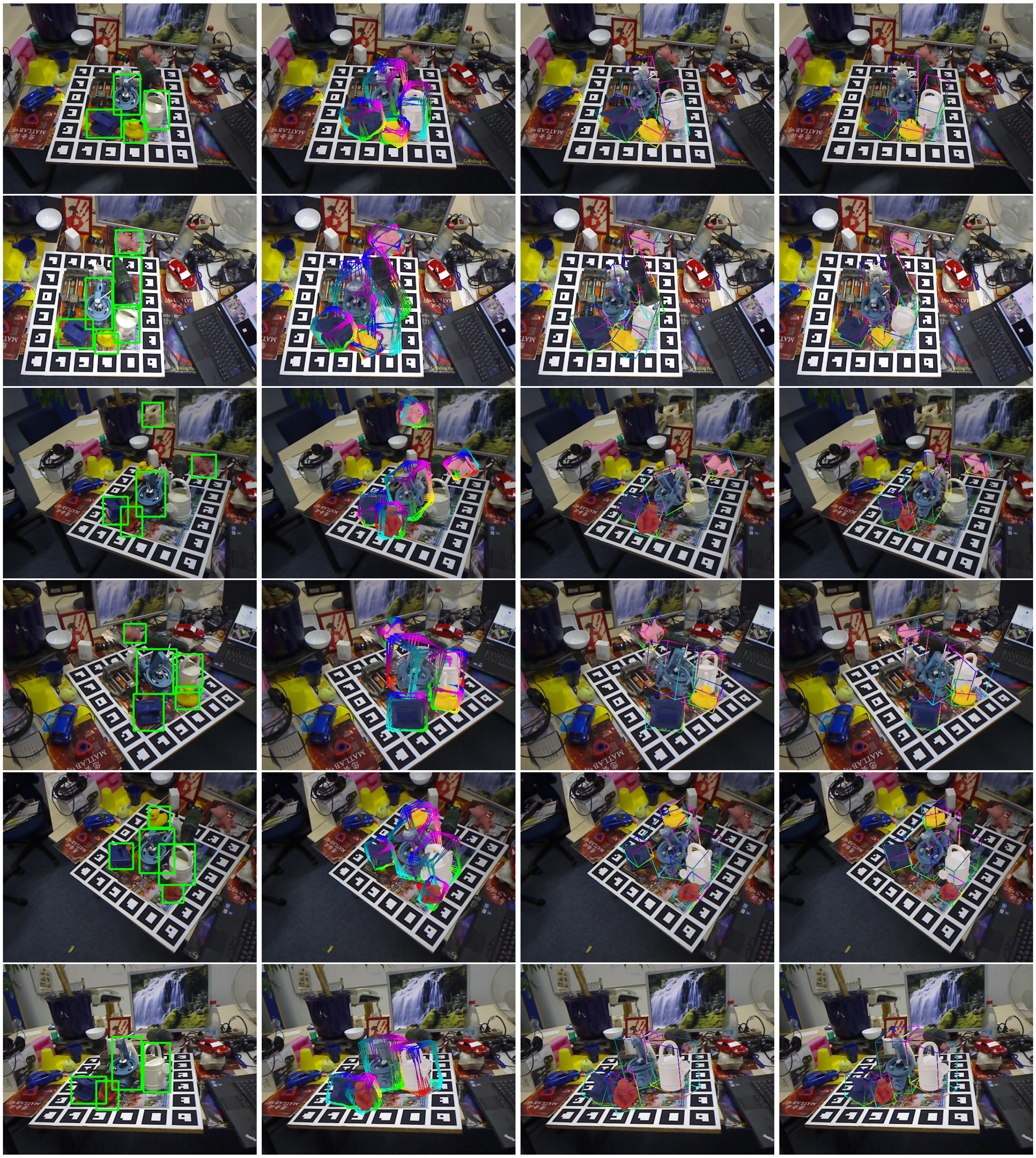


Figure 6: Qualitative results on the multi-object dataset. From left to right: 2D prediction, hypothesis pool, result after 2D refinement, result after 3D refinement.

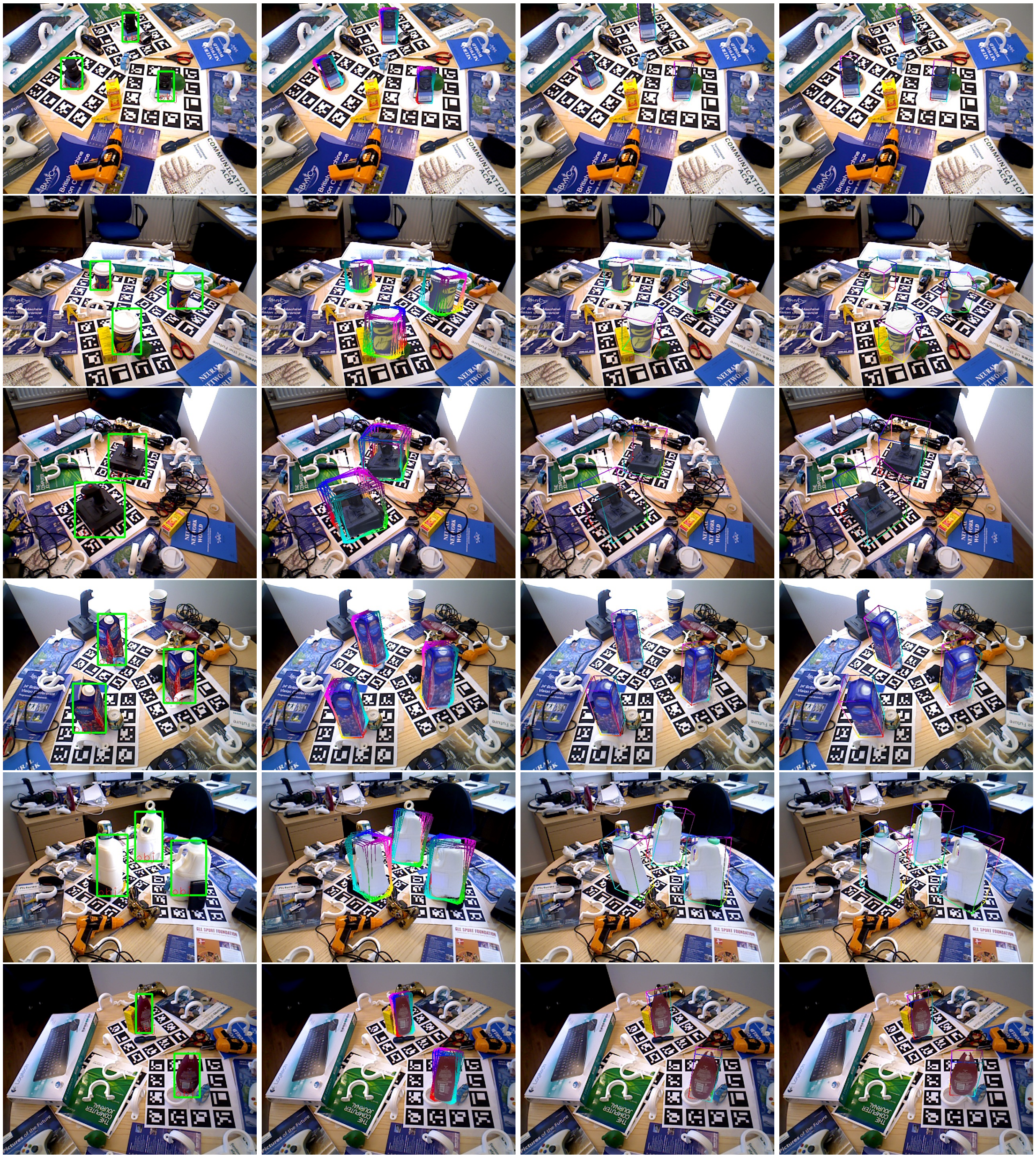


Figure 7: Qualitative results on the Tejani dataset. From left to right: 2D prediction, hypothesis pool, result after 2D refinement, result after 3D refinement.