# Supplementary Material: Depth from Camera Motion and Object Detection

## Depth from Motion Parallax & Detection

Building off of the model from Section 3.1, if there is $x$- or $y$-axis camera motion between observations, we can solve for object dept using corresponding changes in bounding box location (e.g., Box$_1$ to Box$_3$ in Figure 2).

To start, we account for any incidental depth-based changes in scale by reformulating (6) as

$$Z_j = Z_i \Big(\frac{w_i}{w_j}\Big). \tag{21}$$

Next, we use (21) in (4) to relate bounding box center coordinate $x_j$ to corresponding 3D object coordinate $X_j$ as

$$\begin{aligned} x_j &= \frac{f_x X_j}{Z_j} + c_x = \frac{f_x X_j}{Z_i \Big(\frac{w_i}{w_j}\Big)} + c_x \\ \implies & (x_j - c_x)\Big(\frac{w_i}{w_j}\Big) = \frac{f_x X_j}{Z_i}. \end{aligned} \tag{22}$$

Given a static object, changes in lateral object position $X_j$ occur only from changes in camera position $\mathrm{C}_{Xi}$ (2). Thus,

$$X_j - X_i = -(\mathrm{C}_{Xj} - \mathrm{C}_{Xi}). \tag{23}$$

Finally, using (22) and (23), we can solve for $Z_i$ by comparing two observations $i, j$ with motion parallax as

$$\begin{aligned} (x_j - c_x)\Big(\frac{w_i}{w_j}\Big) - (x_i - c_x) &= \frac{f_x(X_j - X_i)}{Z_i} \\ \implies Z_i &= \frac{f_x(\mathrm{C}_{Xi} - \mathrm{C}_{Xj})}{(x_j - c_x)\Big(\frac{w_i}{w_j}\Big) - (x_i - c_x)}. \end{aligned} \tag{24}$$

Notably, (24) can also be derived using vertical motion as

$$Z_i = \frac{f_y(\mathrm{C}_{Yi} - \mathrm{C}_{Yj})}{(y_j - c_y)\Big(\frac{w_i}{w_j}\Big) - (y_i - c_y)}, \tag{25}$$

and scale measure $\frac{h_i}{h_j}$ can replace $\frac{w_i}{w_j}$ in (24) or (25). Also, if there is no $z$-axis camera motion (i.e., $Z_i = Z_j$), then $\frac{w_i}{w_j} = \frac{h_i}{h_j} = 1$ and we can simplify (24) and (25) as

$$Z_i = \frac{f_x(\mathrm{C}_{Xi} - \mathrm{C}_{Xj})}{x_j - x_i} = \frac{f_y(\mathrm{C}_{Yi} - \mathrm{C}_{Yj})}{y_j - y_i}. \tag{26}$$

## Comparison of Depth Estimation Cues

We provide ODMD Test Set results in Table 3 to compare solutions using different depth estimation cues. We evaluate three different analytical solutions that use single cues and DBox$_{\text{NS}}$, which uses full $x, y, z$ motion.

Table 3. **ODMD Results for Various Depth Estimation Cues**.

| | | Mean Percent Error (20) | | | | |
|---|---|---|---|---|---|---|
| | | | Perturb | | | |
| Analytical Depth Estimation Cue | Object Depth Method | Norm. | Camera Motion | Object Detect. | Robot | All Sets |
| Learning-based Methods | | | | | | |
| Full $x, y, z$ Motion | DBox$_{\text{NS}}$ (14) | 0.5 | **3.9** | **6.4** | **12.5** | **5.8** |
| Analytical Methods | | | | | | |
| Optical Expansion | Box$_{\text{LS}}$ (9) | **0.0** | 4.5 | 21.6 | 21.2 | 11.8 |
| Motion Parallax | $Z_n$ (24)-(25) | **0.0** | 33.9 | 51.6 | 65.6 | 37.8 |
| Optical Expansion | $Z_n$ (8) | **0.0** | 5.2 | 80.9 | 124.1 | 52.5 |

For the motion parallax solution, we use the average of the lateral (24) and vertical (25) motion parallax solutions, using scale measure $\frac{w_i}{w_j}$ in (24) and $\frac{h_i}{h_j}$ for (25). Notably, this is a two-observation solution, so we use the end point observations of each example, i.e., $i = n = 10$ and $j = 1$. For comparison, we similarly evaluate a two-observation, optical expansion-based solution, which uses the average of (8) when using $\frac{w_i}{w_j}$ and $\frac{h_i}{h_j}$ for the end point observations.

Motion parallax performs the best overall for the two-observation solutions in Table 3. The optical expansion solution performs surprising well with camera motion perturbations but much worse on the test sets with object detection errors (i.e., Perturb Object Detection and Robot). Both solutions are perfect on the error-free Normal Set.

The Box$_{\text{LS}}$ solution, which uses optical expansion over all $n$ observations, significantly outperforms both two-observation solutions, especially on test sets with object detection errors. Thus, for applications with real-world detection (e.g., the Robot Set), we find that incorporating many observations is more beneficial than choosing between optical expansion or motion parallax with fewer observations.

DBox$_{\text{NS}}$, using all $n$ observations and full $x, y, z$ motion, performs the best overall and on all test sets with any kind of input errors. Admittedly, some error mitigation likely results from DBox$_{\text{NS}}$ using a probabilistic learning-based method. Still, DBox$_{\text{NS}}$ trains on ideal data without any input errors, so DBox$_{\text{NS}}$ predictions are based on an ideal model, just like the analytical methods. Accordingly, we postulate that DBox$_{\text{NS}}$'s improvement over Box$_{\text{LS}}$ is primarily the result of learning full $x, y, z$ motion features, which are more reliable than a single depth cue (e.g., DBox$_p^z$ in Table 1).

Although our analytical model in Section 3.1 and current methods focus on $x, y, z$ camera motion, adding rotation as an additional depth estimation cue is an area of future work. Nonetheless, our state-of-the-art results on the ODMS Driving Set in Table 2 do include examples with camera rotation from vehicle turning [15, Section 5.2] (Figure 6, center). Finally, as a practical consideration for robotics applications, motion planners using our current approach can simply incorporate rotation after estimating depth.

## Camera-based Constraints on Generated Data

When generating new ODMD training data in Section 3.3, we consider the full camera model to ensure that generated 3D objects and their bounding boxes are within the camera's field of view. To derive this constraint, we first note that the center of a bounding box (1) is within view if $x_i \in [0, W_I], y_i \in [0, H_I]$, where $W_I, H_I$ are the image width and height. Using (4), we represent these constraints for 3D camera-frame coordinates $X_i, Y_i$ as

$$
\begin{aligned}
&0 \leq x_i = \frac{f_x X_i}{Z_i} + c_x \leq W_I,\ 0 \leq y_i = \frac{f_y Y_i}{Z_i} + c_y \leq H_I \\
&\qquad \implies \frac{-c_x Z_i}{f_x} \leq X_i \leq \frac{(W_I - c_x) Z_i}{f_x} \qquad (27) \\
&\qquad \implies \frac{-c_y Z_i}{f_y} \leq Y_i \leq \frac{(H_I - c_y) Z_i}{f_y}. \qquad (28)
\end{aligned}
$$

We also consider constraints based on the maximum object size $s_{\text{max}}$ and camera movement range $\Delta \mathbf{p}_{\text{max}}$ (15). We use $\Delta \mathbf{p}_{\text{max}}$ by defining it in terms of its components parts as

$$\Delta \mathbf{p}_{\text{max}} := \left[\Delta \mathrm{C}_{X\text{max}}, \Delta \mathrm{C}_{Y\text{max}}, \Delta \mathrm{C}_{Z\text{max}}\right]^{\intercal}. \qquad (29)$$

Then, using $s_{\text{max}}$, $\Delta \mathbf{p}_{\text{max}}$, and the initial object position $\left[X_1, Y_1, Z_1\right]^{\intercal}$ (16), we update the constraints in (27) as

$$
\begin{aligned}
&\frac{-c_x(Z_1 - \Delta \mathrm{C}_{Z\text{max}})}{f_x} \leq X_1 - \Delta \mathrm{C}_{X\text{max}} - \frac{s_{\text{max}}}{2} \leq \\
&X_1 + \Delta \mathrm{C}_{X\text{max}} + \frac{s_{\text{max}}}{2} \leq \frac{(W_I - c_x)(Z_1 - \Delta \mathrm{C}_{Z\text{max}})}{f_x}, \qquad (30)
\end{aligned}
$$

and, equivalently for height, update constraints in (28) as

$$
\begin{aligned}
&\frac{-c_y(Z_1 - \Delta \mathrm{C}_{Z\text{max}})}{f_y} \leq Y_1 - \Delta \mathrm{C}_{Y\text{max}} - \frac{s_{\text{max}}}{2} \leq \\
&Y_1 + \Delta \mathrm{C}_{Y\text{max}} + \frac{s_{\text{max}}}{2} \leq \frac{(H_I - c_y)(Z_1 - \Delta \mathrm{C}_{Z\text{max}})}{f_y}, \qquad (31)
\end{aligned}
$$

where $Z_1 - \Delta \mathrm{C}_{Z\text{max}}$ accounts for camera approach to the object, $\Delta \mathrm{C}_{X\text{max}}$ and $\Delta \mathrm{C}_{Y\text{max}}$ account for lateral and vertical camera movement, and $\frac{s_{\text{max}}}{2}$ accounts for object width and height. Because (30)-(31) use the maximum camera movement range and object size, they guarantee, first, (27)-(28) are satisfied for all $n$ object positions $\left[X_i, Y_i, Z_i\right]^{\intercal}$ and, second, all corresponding bounding boxes are in view.

Given $X_1, Y_1$, we can find the lower bound for the initial object depth ($Z_{1\text{min}}$) by replacing $Z_1$ with $Z_{1\text{min}}$ in (30) and (31) to find

$$
\begin{aligned}
Z_{1\text{min}} \geq \Delta \mathrm{C}_{Z\text{max}} + \max\Bigg( &\left(\frac{f_x}{c_x}\right)\left(\frac{s_{\text{max}}}{2} + \Delta \mathrm{C}_{X\text{max}} - X_1\right), \\
&\left(\frac{f_y}{c_y}\right)\left(\frac{s_{\text{max}}}{2} + \Delta \mathrm{C}_{Y\text{max}} - Y_1\right), \\
&\left(\frac{f_x}{W_I - c_x}\right)\left(\frac{s_{\text{max}}}{2} + \Delta \mathrm{C}_{X\text{max}} + X_1\right), \\
&\left(\frac{f_y}{H_I - c_y}\right)\left(\frac{s_{\text{max}}}{2} + \Delta \mathrm{C}_{Y\text{max}} + Y_1\right)\Bigg).
\end{aligned}
\qquad (32)
$$

In other words, given an object's center position and maximum size, the minimum viewable depth is constrained by the closest image boundary after camera movement. Note that there is no equivalent upper bound for $Z_{1\text{max}}$.

Given $Z_1$, similar to (32), we can find the lower and upper bounds for $X_1, Y_1$ in (16) using (30) and (31) to find

$$
\begin{aligned}
X_{1\text{min}} &\geq \left(\frac{c_x}{f_x}\right)(\Delta \mathrm{C}_{Z\text{max}} - Z_1) + \Delta \mathrm{C}_{X\text{max}} + \frac{s_{\text{max}}}{2} \\
Y_{1\text{min}} &\geq \left(\frac{c_y}{f_y}\right)(\Delta \mathrm{C}_{Z\text{max}} - Z_1) + \Delta \mathrm{C}_{Y\text{max}} + \frac{s_{\text{max}}}{2} \\
X_{1\text{max}} &\leq \left(\frac{W_I - c_x}{f_x}\right)(Z_1 - \Delta \mathrm{C}_{Z\text{max}}) - \Delta \mathrm{C}_{X\text{max}} - \frac{s_{\text{max}}}{2} \\
Y_{1\text{max}} &\leq \left(\frac{H_I - c_y}{f_y}\right)(Z_1 - \Delta \mathrm{C}_{Z\text{max}}) - \Delta \mathrm{C}_{Y\text{max}} - \frac{s_{\text{max}}}{2}.
\end{aligned}
\qquad (33)
$$

When generating ODMD training data in Section 3.3, we cannot select the $Z_{1\text{min}}$ constraint simultaneously with the $X_{1\text{min}}, Y_{1\text{min}}, X_{1\text{max}}, Y_{1\text{max}}$ constraints in (33). Alternatively, we choose a $Z_{1\text{min}}$ value greater than the lower bound for $X_1, Y_1 = 0$ in (32), then randomly select $Z_1 \sim \mathcal{U}[Z_{1\text{min}}, Z_{1\text{max}}]$ for each training example. Once $Z_1$ is randomly determined, we use (33) to find

$$
\begin{aligned}
X_{1\text{min}}(Z_1) &= \left(\frac{c_x}{f_x}\right)(\Delta \mathrm{C}_{Z\text{max}} - Z_1) + \Delta \mathrm{C}_{X\text{max}} + \frac{s_{\text{max}}}{2} \\
Y_{1\text{min}}(Z_1) &= \left(\frac{c_y}{f_y}\right)(\Delta \mathrm{C}_{Z\text{max}} - Z_1) + \Delta \mathrm{C}_{Y\text{max}} + \frac{s_{\text{max}}}{2} \\
X_{1\text{max}}(Z_1) &= \left(\frac{W_I - c_x}{f_x}\right)(Z_1 - \Delta \mathrm{C}_{Z\text{max}}) - \Delta \mathrm{C}_{X\text{max}} \\
&\qquad - \frac{s_{\text{max}}}{2} \\
Y_{1\text{max}}(Z_1) &= \left(\frac{H_I - c_y}{f_y}\right)(Z_1 - \Delta \mathrm{C}_{Z\text{max}}) - \Delta \mathrm{C}_{Y\text{max}} \\
&\qquad - \frac{s_{\text{max}}}{2}, \qquad (34)
\end{aligned}
$$

which is the exact solution we use in (16). Notably, in absence of making adjustments for the specific object size or camera movement range of each example, (34) provides the

Table 4. **Detailed ODMD Results**.

| Object | Percent Error (20) | | | | |
|---|---|---|---|---|---|
| Depth | | | Range | | Standard |
| Method | Mean | Median | Minimum | Maximum | Deviation |
| | | | Normal Set | | |
| DBox$_p$ | 1.73 | 0.96 | 0.0002 | 48.21 | 2.90 |
| DBox$_{\text{Abs}}$ | 1.11 | 0.82 | 0.0004 | 21.10 | 1.19 |
| DBox$_{\text{NS}}$ | 0.54 | 0.38 | 0.0001 | 8.68 | 0.63 |
| Box$_{\text{LS}}$ | **0.00** | **0.00** | **0.0000** | **0.00** | **0.00** |
| DBox$^z_p$ | 12.89 | 8.54 | 0.0062 | 80.74 | 13.23 |
| | | | Perturb Camera Motion Set | | |
| DBox$_p$ | 2.45 | 1.86 | 0.0008 | 23.61 | 2.28 |
| DBox$_{\text{Abs}}$ | **2.05** | **1.55** | **0.0002** | **19.45** | **1.96** |
| DBox$_{\text{NS}}$ | 3.91 | 2.93 | 0.0021 | 47.94 | 3.82 |
| Box$_{\text{LS}}$ | 4.47 | 3.13 | 0.0007 | 43.02 | 4.57 |
| DBox$^z_p$ | 12.48 | 8.42 | 0.0025 | 74.18 | 12.18 |
| | | | Perturb Object Detection Set | | |
| DBox$_p$ | 2.54 | 1.54 | 0.0020 | 45.94 | 3.39 |
| DBox$_{\text{Abs}}$ | **1.75** | **1.26** | 0.0007 | **19.51** | **1.81** |
| DBox$_{\text{NS}}$ | 6.35 | 1.98 | 0.0005 | 415.68 | 19.56 |
| Box$_{\text{LS}}$ | 21.60 | 8.90 | **0.0003** | 158.04 | 28.27 |
| DBox$^z_p$ | 15.00 | 9.83 | 0.0189 | 296.31 | 16.93 |
| | | | Robot Set | | |
| DBox$_p$ | **11.17** | 8.31 | 0.0022 | 253.02 | **13.94** |
| DBox$_{\text{Abs}}$ | 13.29 | 9.44 | 0.0024 | **223.76** | 14.90 |
| DBox$_{\text{NS}}$ | 12.47 | **8.11** | 0.0092 | 656.85 | 25.03 |
| Box$_{\text{LS}}$ | 21.23 | 12.17 | **0.0010** | 262.48 | 26.92 |
| DBox$^z_p$ | 21.96 | 14.64 | 0.0099 | 342.40 | 26.39 |

Table 5. **Detailed ODMS Results**.

| Object | Percent Error (20) | | | | |
|---|---|---|---|---|---|
| Depth | | | Range | | Standard |
| Method | Mean | Median | Minimum | Maximum | Deviation |
| | | | Normal Set | | |
| DBox$^z_p$ | 11.82 | 8.17 | 0.0049 | 167.80 | 12.42 |
| Box$_{\text{LS}}$ | 13.66 | 10.49 | **0.0025** | **137.39** | 11.97 |
| DBox$^z_{\text{NS}}$ | **9.20** | **6.69** | 0.0048 | 146.79 | **9.55** |
| DBox$^z_{\text{Abs}}$ | 21.31 | 11.98 | 0.0119 | 451.54 | 33.95 |
| | | | Perturb Set | | |
| DBox$^z_p$ | **20.34** | 15.25 | **0.0008** | 220.46 | **19.73** |
| Box$_{\text{LS}}$ | 36.62 | 27.76 | 0.0050 | **141.85** | 30.06 |
| DBox$^z_{\text{NS}}$ | 31.55 | 19.95 | 0.0205 | 644.55 | 48.03 |
| DBox$^z_{\text{Abs}}$ | 25.49 | **15.12** | 0.0033 | 265.11 | 30.68 |
| | | | Robot Set | | |
| DBox$^z_p$ | **11.45** | 6.29 | 0.0061 | 418.41 | **23.81** |
| Box$_{\text{LS}}$ | 17.62 | 9.15 | **0.0011** | 390.12 | 34.22 |
| DBox$^z_{\text{NS}}$ | 39.25 | **5.97** | 0.0082 | 8778.45 | 310.94 |
| DBox$^z_{\text{Abs}}$ | 20.36 | 10.28 | 0.0033 | **358.86** | 32.80 |
| | | | Driving Set | | |
| DBox$^z_p$ | **24.84** | **18.99** | 0.0323 | **213.93** | **22.83** |
| Box$_{\text{LS}}$ | 33.29 | 26.50 | 0.1783 | 294.91 | 31.10 |
| DBox$^z_{\text{NS}}$ | 37.31 | 21.43 | **0.0108** | 613.14 | 55.75 |
| DBox$^z_{\text{Abs}}$ | 53.13 | 55.89 | 0.0878 | 296.88 | 26.65 |

Table 6. **Detailed ODMD Results (Absolute Error)**.

| Object | Absolute Error (35) | | | | |
|---|---|---|---|---|---|
| Depth | | | Range | | Standard |
| Method | Mean | Median | Minimum | Maximum | Deviation |
| | | | Normal Set (cm) | | |
| DBox$_p$ | 1.42 | 0.69 | 0.0002 | 46.57 | 2.62 |
| DBox$_{\text{Abs}}$ | 0.87 | 0.59 | 0.0003 | 11.56 | 0.95 |
| DBox$_{\text{NS}}$ | 0.41 | 0.28 | 0.0001 | 8.97 | 0.53 |
| Box$_{\text{LS}}$ | **0.00** | **0.00** | **0.0000** | **0.00** | **0.00** |
| DBox$^z_p$ | 10.30 | 6.22 | 0.0040 | 77.13 | 11.68 |
| | | | Perturb Camera Motion Set (cm) | | |
| DBox$_p$ | 1.93 | 1.38 | 0.0005 | 20.84 | 1.93 |
| DBox$_{\text{Abs}}$ | **1.63** | **1.13** | **0.0001** | **17.17** | **1.65** |
| DBox$_{\text{NS}}$ | 3.04 | 2.16 | 0.0024 | 41.03 | 3.12 |
| Box$_{\text{LS}}$ | 3.44 | 2.37 | 0.0005 | 33.65 | 3.66 |
| DBox$^z_p$ | 9.92 | 6.26 | 0.0022 | 67.80 | 10.57 |
| | | | Perturb Object Detection Set (cm) | | |
| DBox$_p$ | 2.06 | 1.11 | 0.0017 | 42.25 | 3.07 |
| DBox$_{\text{Abs}}$ | **1.39** | **0.93** | 0.0005 | **15.91** | **1.55** |
| DBox$_{\text{NS}}$ | 5.01 | 1.47 | 0.0004 | 281.55 | 15.07 |
| Box$_{\text{LS}}$ | 17.58 | 7.08 | **0.0002** | 121.45 | 23.67 |
| DBox$^z_p$ | 11.81 | 7.12 | 0.0089 | 146.10 | 13.43 |
| | | | Robot Set (cm) | | |
| DBox$_p$ | **8.08** | 5.79 | 0.0012 | 260.28 | 12.06 |
| DBox$_{\text{Abs}}$ | 8.83 | 6.71 | 0.0018 | **55.83** | **7.87** |
| DBox$_{\text{NS}}$ | 9.23 | **5.57** | 0.0045 | 579.77 | 23.51 |
| Box$_{\text{LS}}$ | 14.49 | 8.63 | **0.0007** | 197.56 | 17.70 |
| DBox$^z_p$ | 14.65 | 10.29 | 0.0089 | 161.98 | 14.69 |

Table 7. **Detailed ODMS Results (Absolute Error)**.

| Object | Absolute Error (35) | | | | |
|---|---|---|---|---|---|
| Depth | | | Range | | Standard |
| Method | Mean | Median | Minimum | Maximum | Deviation |
| | | | Normal Set (cm) | | |
| DBox$^z_p$ | 3.65 | 2.70 | 0.0020 | **30.77** | **3.66** |
| Box$_{\text{LS}}$ | 4.74 | 3.43 | **0.0006** | 55.77 | 4.81 |
| DBox$^z_{\text{NS}}$ | **2.98** | **2.14** | 0.0008 | 82.58 | 3.74 |
| DBox$^z_{\text{Abs}}$ | 5.57 | 3.98 | 0.0059 | 50.57 | 5.81 |
| | | | Perturb Set (cm) | | |
| DBox$^z_p$ | 7.19 | **4.46** | **0.0003** | 77.16 | 8.06 |
| Box$_{\text{LS}}$ | 15.17 | 8.30 | 0.0016 | 79.01 | 16.45 |
| DBox$^z_{\text{NS}}$ | 12.21 | 5.77 | 0.0091 | 295.98 | 23.27 |
| DBox$^z_{\text{Abs}}$ | **6.68** | 5.20 | 0.0004 | **37.75** | **5.65** |
| | | | Robot Set (cm) | | |
| DBox$^z_p$ | **3.34** | 1.78 | 0.0013 | 88.89 | **5.94** |
| Box$_{\text{LS}}$ | 5.21 | 2.58 | **0.0005** | 79.71 | 10.54 |
| DBox$^z_{\text{NS}}$ | 12.06 | **1.71** | 0.0019 | 1634.35 | 84.61 |
| DBox$^z_{\text{Abs}}$ | 5.64 | 3.04 | 0.0010 | **70.20** | 7.93 |
| | | | Driving Set (m) | | |
| DBox$^z_p$ | **3.63** | **1.86** | 0.0031 | **37.95** | **5.00** |
| Box$_{\text{LS}}$ | 5.08 | 2.35 | 0.0142 | 58.07 | 8.07 |
| DBox$^z_{\text{NS}}$ | 5.03 | 2.37 | **0.0004** | 105.97 | 8.43 |
| DBox$^z_{\text{Abs}}$ | 9.05 | 5.82 | 0.0046 | 57.60 | 9.24 |

greatest range of initial positions that also guarantees the object is in view for all $n$ observations. Finally, (34) is linear, so we vectorize it for large batches of training examples.

## Detailed ODMD and ODMS Results

We provide more comprehensive and detailed ODMD and ODMS results in Tables 4 and 5. Specifically, we provide a more precise mean percent error (20) and include the median, range, and standard deviation for each test set.

We also provide ODMD and ODMS results for the absolute error in Tables 6 and 7, which we calculate for each example as

$$\text{Absolute Error} = \left| Z_n - \hat{Z}_n \right|, \tag{35}$$

where $Z_n$ and $\hat{Z}_n$ are ground truth and predicted object depth at final camera position $\mathbf{p}_n$. Notably, we use percent

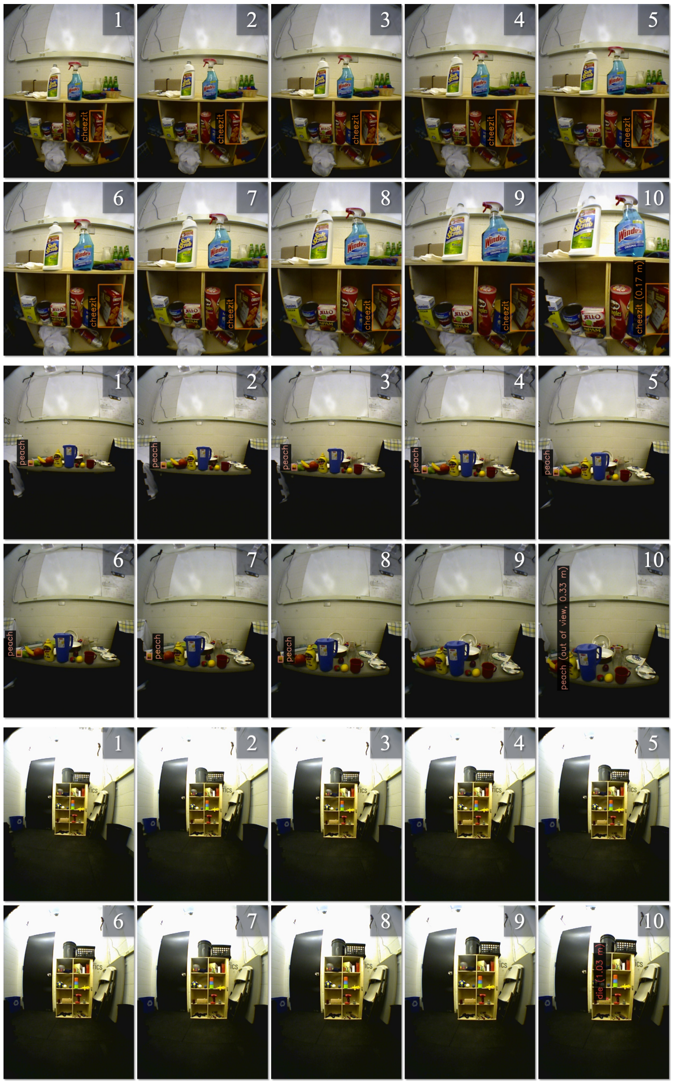

Figure 8. **ODMD Robot Test Set Examples** (best viewed in electronic format). Every two rows show a ten-observation example progressing from left to right, and we show the ground truth object depth in the final image. In the cheezit example (top two rows), the camera perspective changes and the detected object partially leaves view (observations 8-10), causing a distortion to the bounding box shape relative to earlier observations. In the peach example (middle two rows), the detected object completely leaves view during the final two observations (9-10), providing no bounding box information at the prediction location. Finally, for the 16 mm die example (bottom), the camera starts far away from the small object (1), which is not detected until the camera is closer in the final observation (10).

error (20) in the paper to provide a consistent comparison across domains (and examples) with markedly different object depth distances. For example, the 0.10 m absolute error from Figure 7 is a much better result for a camera phone application than it would be for robot grasping.

## Object Motion Considerations

The ODMS Driving Set includes moving objects [[15], Section 5.2]. On the other hand, our analytical model in Section 3.1 assumes static objects. Nonetheless, $\text{DBox}_p^z$ achieves the current state-of-the-art result on the ODMS Driving Set in Table 2. We attribute $\text{DBox}_p^z$'s success to training with camera movement perturbations (18). Note that training with these perturbations improves robustness to input errors for the relative distance changes between the camera and object, whether caused by camera motion errors *or* unintended motion of the object itself. In general, objects that move much less than the camera are not an issue.

## ODMD Robot Test Set Examples

For the ODMD Robot Test Set, we intentionally select challenging objects and settings that make object detection and depth estimation difficult. To illustrate this point, we show a few example challenges in Figure 8.