We present a new, publicly-available image dataset generated by the NVIDIA Deep Learning Data Synthesizer intended for use in object detection, pose estimation, and tracking applications. This dataset contains 144k stereo image pairs that synthetically combine 18 camera viewpoints of three photorealistic virtual environments with up to 10 objects (chosen randomly from the 21 object models of the YCB dataset) and flying distractors. Object and camera pose, scene lighting, and quantity of objects and distractors were randomized. Each provided view includes RGB, depth, segmentation, and surface normal images, all pixel level.

We describe our approach for domain randomization and provide insight into the decisions that produced the dataset.

1. Introduction

Computer vision has shifted towards the use of neural networks, trained on large datasets, to exceed the performance of more traditional algorithms in object recognition, pose estimation, and object tracking. In order for training to be effective, the training set needs to exhibit similar characteristics to the target application, while maintaining enough variation to prevent overfitting. While a number of datasets to support object recognition research already exist, which provide a variety of labels, there is still a need for larger, high-quality datasets. To this end, we present a new synthetic dataset called SIDOD as an acronym for Synthetic Image Dataset for 3D Object Pose Recognition with Distractors present. We are motivated by the notion that object pose and segmentation are much easier to capture with synthetic tools rather than custom-built sensors. An added benefit is other potentially useful or interesting pieces of information, such as surface normals, are available in the rendering process used to generate synthetic images.

Given that SIDOD provides 3D pose and rotation along with depth and pixel-level segmentation, it may be useful for a number of applications in robotics and computer vision. We were motivated by potential uses in Virtual and Augmented Reality (VR/AR) where object understanding and tracking can be vital to immersive user experiences. One challenge in tracking hand-held objects is the necessary occlusion of the user’s hand, over a range of grasps, as well as a wide range of possible colors and textures. To provide random occlusion to approximate these anticipated situations, we added subsets of images with flying distractors to our dataset. Flying distractors are 3D shapes, which are not objects of interest.

Similarly to the FALLING THINGS dataset, we provide stereo views as RGB, depth, and pixel-level segmentation images and increase the amount of available data with significant object occlusion. Additionally, we include flying distractors in a subset of views, representing even more aggressive occlusion conditions (Fig. 2). We also provide pixel-level surface normals for all view pairs. A comparison with existing datasets is given in Table 1.

2. SIDOD Dataset

We generated our dataset using the NVIDIA Deep Learning Data Synthesizer (NDDS). Built on top of Unreal Engine 4.18.3, this tool captures view data at a much higher rate than many other available rendering solutions, as high as tens of frames per second. We selected three distinct virtual environments and manually chose a variety of
distinct locations within each environment to provide sufficient variations of viewpoints to mimic a variety of indoor locations. The scenes are as follows: bakery (4 locations), AIUEV02_001 (10 locations), AIUEV02_002 (4 locations) for a total of 18 different locations. Corresponding directories can be found for each location. The scenes prefixed with AIUE\(^2\) represent common, indoor, residential home settings, while the bakery scene represents more of a typical commercial business. We use the 21 objects from YCB [1] that are included in the FALLING THINGS dataset.

For each stereo frame captured, either a single object was selected at random, or 4 to 10 objects were drawn. Parameters we randomized include object pose (random rotation and placement within a cylinder of 60 radius and 5 height units), camera pose, scene lighting (hue, intensity, additional point sources), quantity of objects for mixed captures, and quantity of flying distractors (where present). Similar to FALLING THINGS, the objects of interest follow physics simulation with gravity enabled. The lighting of the scene comes from a point light that is placed 50cm above each scene point with brightness varying from 5,000–20,000 units and a randomized color hue. When flying distractors are enabled, we select 50–100 from 10 possible meshes (e.g., cubes, cylinders, cones) which appear at random locations with random levels of visibility. Each flying distractor uses one of 7 available material textures at random.

We provide labels as JSON files (two per stereoscopic view pair) including information such as camera location/pose and object location, pose, visibility, and bounding boxes (world space/screen space). Labels are affiliated to their matching images using their filename prefix. The dataset is broken down into captures containing single objects and multiple (mixed) objects, as well as with and without distractors. For each of these 4 variations, we capture 2,000 stereo views from each of 18 different viewpoints across 3 scenes (possibly with different camera orientation) for a total of 144k stereo captures, or 288k individual views if stereo pairs are considered separately. Unlike the FALLING THINGS dataset, we did not generate multiple frames during a falling animation, and instead provide individual frames with the randomized parameters and objects.

**Statistical analysis:** Figure 3 displays the total number of occurrences of each object across our entire dataset. Note that the results provide a variety of high-visibility and low-visibility cases where object occlusion is present. We hope these occlusions can provide meaningful data for training systems to deal with user hand occlusions present during interaction. The amount of occlusion varies based on the relative size of the objects and the other objects and distractors that may also be present. Figure 4 shows the pitch/yaw/roll angles, distance to/from the camera, and screen location heatmap for one object in the 21 chosen from the YCB dataset (mustard bottle).

One additional consideration is that the flying distractors in particular were not carefully controlled to avoid clipping through other objects in the scene. These objects are not intended to directly represent objects that might exist in the real world, and instead provide noise and variation similar to unexpected objects, therefore this clipping should not significantly affect their usefulness beyond the already non-

---

Table 1. Object pose and detection datasets with at least 10,000 frames.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#objects</th>
<th>#frames</th>
<th>Description</th>
<th>depth</th>
<th>stereo</th>
<th>3D pose</th>
<th>full rotation</th>
<th>occlusion</th>
<th>extreme lighting</th>
<th>segmentation</th>
<th>bbox coords</th>
<th>flying distractors</th>
</tr>
</thead>
<tbody>
<tr>
<td>UW RGB [4]</td>
<td>300</td>
<td>250k</td>
<td>various</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>LINEMOD [2]</td>
<td>15</td>
<td>18k</td>
<td>various</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Pascal3D+ [8]</td>
<td>12</td>
<td>30k</td>
<td>various</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Rutgers APC [5]</td>
<td>24</td>
<td>10k</td>
<td>warehouse</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>T-LESS [3]</td>
<td>30</td>
<td>10k</td>
<td>industrial</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>YCB Video [9]</td>
<td>21</td>
<td>134k</td>
<td>household</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Falling Things [7]</td>
<td>21</td>
<td>60k</td>
<td>household</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SIDOD</td>
<td>21</td>
<td>144k</td>
<td>household</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

---

https://evermotion.org

---
realistic inclusion of them in the first place. It may therefore make sense to use different learning rates when training a network on the part of the dataset with flying distractors than should be used with the rest of the dataset.

3. Conclusion

We present a new dataset with distractors created using the NVIDIA NDDS toolkit [6]. This dataset includes 18 view locations from 3 scenes and includes surface normals, flying distractors, and extreme lighting variations. SIDOD is more than twice as large as the FALLING THINGS dataset and provides significantly more parameter variations. We hope that this dataset will enable researchers in the areas of computer vision and robotics to further their goals in training networks more robust to occlusion and unexpected environmental factors.

4. Acknowledgements

We would like to acknowledge the many discussions we had during the development of this dataset with our collaborators at NVIDIA including Niveditha Kalavakonda, Turner Whitted, David Luebke, Thang To, Jonathan Tremblay, Omer Shapira, and others. M.B. acknowledges support by DARPA SBIR/STTR.

References