https://doi.org/10.2352/ISSN.2470-1173.2020.8.IMAWM-221

Extra FAT: A Photorealistic Dataset for 6D Object Pose Estimation *

Jianhang Chen[1], Daniel Mas Montserrat[1], Qian Lin[2], Edward J. Delp[1], Jan P. Allebach[1];
[1]School of Electrical and Computer Engineering, Purdue University; West Lafayette, Indiana, USA
[2]HP Labs, HP Inc; Palo Alto, California, USA

Abstract

We introduce a new image dataset for object detection and 6D pose estimation, named Extra FAT. The dataset consists of 825K photorealistic RGB images with annotations of ground-truth location and rotation for both the virtual camera and the objects. A registered pixel-level object segmentation mask is also provided for object detection and segmentation tasks. The dataset includes 110 different 3D object models. The object models were rendered in five scenes with diverse illumination, reflection, and occlusion conditions.

Introduction

Pose estimation of surrounding objects serves as the basis of various computer vision applications such as virtual reality (VR), augmented reality (AR), robotic manipulation, autonomous navigation, and human-machine interaction. For example, to insert a virtual object in an AR application, it should be accurately registered with the real world. In order to do so, the geometry, pose, and shape of the objects and surfaces composing a scene need to be inferred from the image, video and/or depth information. For tasks such as autonomous navigation or robot manipulation, the pose of the objects needs to be estimated in order to properly move the robot or vehicle. In order to understand the geometry and position of the objects composing a scene, object detection and pose estimation techniques are required.

Traditionally, such methods have used RGB-D images in order to infer the pose of the objects. The main drawback of such an approach is that depth cameras are not widely available (e.g. smartphones) and typically have low resolution and low frame rate making it difficult to detect very small, thin, or fast-moving objects. Therefore, RGB-only based methods are preferred. Recently, many methods based on deep learning have been presented. These methods use convolutional neural networks to estimate the 6D pose of objects. Such neural networks estimate the pose by detecting keypoints [13], estimating a 3 dimensional bounding box [19, 7, 15], matching the input image with rendered images [10, 12], or directly treating pose estimation as a classification [9] or regression [20] problem.

With the increasing number of deep learning based methods for RGB-only pose estimation, there is a need for more training data. Capturing real images is highly time-consuming, therefore a faster approach is preferred. In addition, annotating the pose of objects manually is tedious and inaccurate. While several image synthesis methods have been presented [11] to automatically generate new training samples, the resulting images can lack realistic appearance. An efficient and effective alternative is photorealistic image rendering. Photorealistic rendering allows easy generation of a large number of images containing realistic lighting, occlusions, and real-world distortions with ground truth labeled automatically.

There are many publicly available datasets consisting of real-world images for 6D pose estimation. For example, T-LESS [5] is a dataset with 30 industrial objects that lack distinctive texture. There are 48.9K images in the T-LESS dataset. Many objects in T-LESS dataset are symmetric; and the similarity among them is challenging for pose estimation task. The YCB dataset [2] contains 9.24K images of 77 real-life objects for benchmarking in robot grasping and manipulation tasks. The images in the YCB dataset are captured by the BigBIRD Object Scanning Rig and the Google scanner. The YCB-Video [20] dataset has 134K video frames for 21 household objects taken from the YCB dataset. The LINEMOD dataset [4] is another widely used public dataset for 6D pose estimation with various toys and household objects. The LINEMOD OCCLUSION dataset [1, 8] is a complementary dataset for the LINEMOD dataset with 10K images under different lighting and occlusion conditions. The Rutgers APC [16] dataset includes real images of textured products used in the first Amazon Picking Challenge. The images in the Rutgers APC dataset with different poses and clutter conditions are mainly used for training algorithms in warehouse objects pick-and-place. The IC-MI dataset proposed in [17] has images for six objects heavily 2D and 3D cluttered with foreground occlusion.

Due to the large variety of datasets and evaluation metrics and the lack of a common benchmark procedure, the BOP dataset [6] was introduced. The BOP dataset has a thorough survey of 8 different datasets containing images and the evaluation methodologies. Additionally, the TUD Light dataset and TOYOTA Light dataset are introduced in the BOP dataset. The TUD Light dataset includes images of three objects without occlusion under different illuminations. The TOYOTA Light dataset has 21 objects in total. Each object in TOYOTA Light is put on top of a table with different tablecloths and five different lighting conditions. The MVTec Industrial 3D Object Detection Dataset (MVTec ITODD) [3] contains 28 industrial objects. The dataset focuses more on practical and challenging tasks such as industrial bin picking and 3D object inspection. Besides datasets containing real captured images, some photorealistic rendered datasets, such as Falling Things (FAT) [18], have been made publicly available. The FAT dataset contains synthetic images with the 21 household object models from the YCB dataset.

In this paper, we introduce a new dataset named Extra FAT. We follow a similar approach as in the FAT dataset [18]; but we

*Research supported by HP Labs, Inc., Palo Alto, CA 94304.

Table 1: Comparison of different 3D datasets: LINEMOD dataset [4] (LM), YCB dataset [2] (YCB), T-LESS [5] (T-LESS), IC-MI dataset [17] (IC-MI), TOYOTA Light dataset [6] (TYO-L), Rutgers APC dataset [16] (RU-APC), and TUD Light dataset [6] (TUD-L)

Dataset	# obj	# frames	Type	LM	YCB	T-LESS	IC-MI	TYO-L	RU-APC	TUD-L
LINEMOD [4]	15	18K	real	✓						
LM OCC [1, 8]	15	18K	real	✓						
YCB-Video [20]	21	134K	real		✓					
FAT [18]	21	60K	rendered		✓					
T-LESS [5]	30	48.9K	real			✓				
IC-MI [17]	6	4.2K	real				✓			
TYO-L [6]	21	55.4K	combined					✓		
RU-APC [16]	24	10K	real						✓	
TUD-L [6]	6	62.3K	combined							✓
BOP [6]	89	>294K	combined	✓		✓	✓	✓	✓	✓
Extra FAT (ours)	110	825K	rendered	✓	✓	✓	✓	✓	✓	✓

include a larger number of object models and a larger variety of virtual scenes. This dataset includes rendered images containing a large number of 3D object models from the most commonly used datasets for 6D pose estimation. Table 1 compares our dataset with previously presented datasets [†].

For each rendered image in the Extra FAT dataset, the location and rotation for both the virtual camera and the objects, and a registered pixel-level object segmentation mask with 640 × 480 resolution, shown in Figure 1, are provided. Such images and annotations can be used to train and test methods for object detection, segmentation, and pose estimation.

The images are simulated in five different indoor scenes with various illumination and occlusion conditions, as shown in Figure 2. The indoor scenes include common environments such as office spaces, living rooms, and kitchens. There are 825K images in total. The specifications for the Extra FAT dataset are shown in Table 2.

Table 2: Dataset Specification.

Extra FAT Dataset	
Image Resolution	640 × 480
Field of view	90°
Number of frames	825K
Number of objects	110
Number of scenes	5

Dataset

Image Generation

The Extra FAT dataset is generated by rendering 110 3D object models: 21 household objects taken from the publicly available YCB dataset, 15 objects from the LINEMOD dataset, 30 objects from the T-LESS dataset, 14 objects from the Amazon Picking Challenge 2015 dataset, 6 objects from the IC-MI dataset, 3 objects from the TUD Light dataset and 21 objects from the TOYOTA Light dataset, as shown in Figure 3, Figure 4 and Figure 5.

As in the FAT dataset, we use the Unreal Engine 4 (UE4) [14], a commonly used tool for game development, to render 3D object models in the virtual game scenery. The open-source UnrealCV [14] plugin serves as a communication tool to generate photorealistic images and pose ground truth.

In the FAT dataset, objects are placed at random positions from where they fall. In the Extra FAT dataset, we move the object between pre-defined points (therefore our objects are not technically falling but are flying).

We first manually specify some candidate points within the virtual scene. During the image generation process, pairs of candidate points are selected randomly and the virtual camera and object trajectories are defined by linear interpolation between the two points, as shown in Figure 6. While moving the objects between the pair of points, we apply a uniform random perturbation in the location and rotation of the object and the virtual camera.

Statistics of the objects in the Extra FAT dataset show that the distributions of the Yaw, Pitch, and Roll angles are uniform, which indicates that the poses of objects in the dataset are comprehensive and representative for the general pose estimation task.

In order to avoid placing the object out of the visible range of the camera, we constrain the relative location and rotation between the camera and objects. As shown in Figure 7, the pixel coordinates (p_x, p_y) obey the following constraint:

$$\begin{aligned} -\mu_x < p_x < \mu_x \\ -\mu_y < p_y < \mu_y \end{aligned} \tag{1}$$

where $\mu_x = 200$ and $\mu_y = 180$.

The constraint of the relative object position with respect to the camera (t_x, t_y, t_z) can be computed from the pixel coordinates:

$$\begin{aligned} t_x = p_x \frac{t_z}{f_x} \\ t_y = p_y \frac{t_z}{f_y} \end{aligned} \tag{2}$$

where f_x, f_y are the focal lengths in the x and y direction. The parameter t_z is in the range $(\theta_{z1}, \theta_{z2})$ to make sure that the object is in front of the camera and not too far from it, or too close to it. We set $\theta_{z1} = 0.3$ and $\theta_{z2} = 0.8$.

[†] The number of objects in the BOP dataset is from the BOP benchmark paper [6]. There are more models provided for the BOP 2019 challenge.

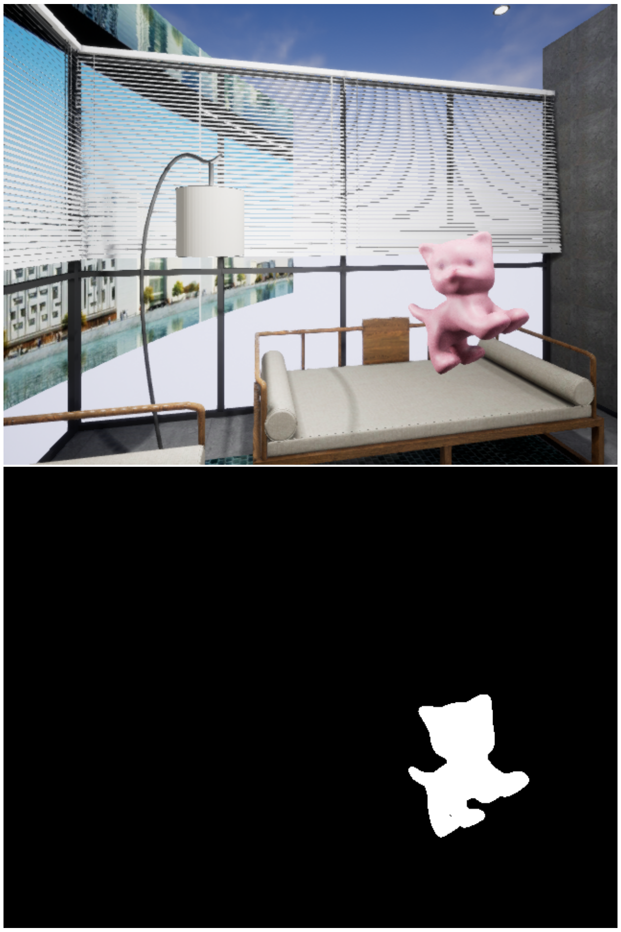


Figure 1. *Each frame in the Extra FAT dataset consists of an image with* 640×480 *resolution, a registered pixel-level object segmentation mask, and the pose ground truth of the virtual camera and the objects*

In order to avoid having objects highly occluded by a wall or other objects in the scenery, we add a constraint on the ratio of mask area to image size:

$$\frac{\sum_{mask} 1}{w \times h} > threshold \tag{3}$$

Images where the segmentation mask area to image size ratio is lower than $threshold = 0.05$ are discarded.

Training and Testing Setting

We propose three different training/testing split approaches. First, we provide a training/testing split with about 6,000 frames for training and 1,500 for testing for each object. Second, the training and testing sets can be divided by scene. Four scenes can be used for training and the other one for testing. Finally, we propose using Extra FAT entirely as a training set and use the BOP [6] benchmark as a testing method.

Conclusion

In this paper, we presented a new dataset for 6D object pose estimation. By using photorealistic rendering, we obtain images with diversity in terms of illumination, reflection, and occlusion. These images can be used to train convolutional neural networks for object detection, segmentation, and pose estimation. We hope Extra FAT will help the community to propose novel algorithms for RGB-only image object segmentation and 6D pose estimation. We invite other researchers to generate new datasets including objects from other commonly used datasets.

References

[1] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton,and C. Rother, Learning 6D Object Pose Estimation using 3D Object Coordinates, Proceedings of the European Conference on Computer Vision, pages 536–551 (2014)

[2] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar, Yale-CMU-Berkeley Dataset for Robotic Manipulation Research, International Journal of Robotics Research, 36(3), 261–268 (2017)

[3] B. Drost, M. Ulrich, P. Bergmann, P. Hartinger, and C. Ste-ger, Introducing MVTec ITODD - a Dataset for 3D Object Recognition in Industry, Proceedings of the IEEE International Conference on Computer Vision, pages 2200–2208 (2017)

[4] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski,K. Konolige, and N. Navab, Model Based Training, Detection and Pose Estimation of Texture-less 3D Objects in Heavily Cluttered Scenes, Proceedings of the Asian Conference on Computer Vision, pages 548–562, (2012)

[5] T. Hodan, P. Haluza, Š Obdržálek, J. Matas, M. Lourakis, and X.s Zabulis, T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects, Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pages 880–888 (2017)

[6] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. GlentBuch,D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, et al, BOP: Benchmark for 6D Object Pose Estimation, Proceedings of the European Conference on Computer Vision, pages 19–34 (2018)

[7] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, SSD-6D: Making RGB-based 3D Detection and 6D Pose Estimation Great Again, Proceedings of the IEEE International Conference on Computer Vision, pages 1530–1538 (2017)

[8] A. Krull, E. Brachmann, F. Michel, M. Ying Yang, S. Gumhold, and C. Rother, Learning Analysis by Synthesis for 6D Pose Estimation in RGB-D Images, Proceedings of the IEEE International Conference on Computer Vision, pages 954–962 (2015)

[9] A. Kundu, Y. Li, and J. M. Rehg, 3D-RCNN: Instance Level 3D Object Reconstruction via Render-and-compare, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3559–3568 (2018)

[10] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, DeepIM: Deep Iterative Matching for 6D Pose Estimation, Proceedings of the European Conference on Computer Vision, pages 683–698 (2018)

[11] D. Mas Montserrat, Q. Lin, J. Allebach, and E. J. Delp, Logo Detection and Recognition with Synthetic Images, Proceedings of the IS&T International Symposium on Electronic Imaging (2018)

[12] D. M. Montserrat, J. Chen, Q. Lin, J. P. Allebach, and E. J.Delp, Multi-View matching Network for 6D Pose Estimation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2019)

[13] G. Pavlakos, X. Zhou, A. Chan, K. G. Derpanis, and K. Daniilidis, 6-DoF Object Pose from Semantic Keypoints, Proceedings of the International Conference on Robotics and Automation, pages 2011–2018 (2017)

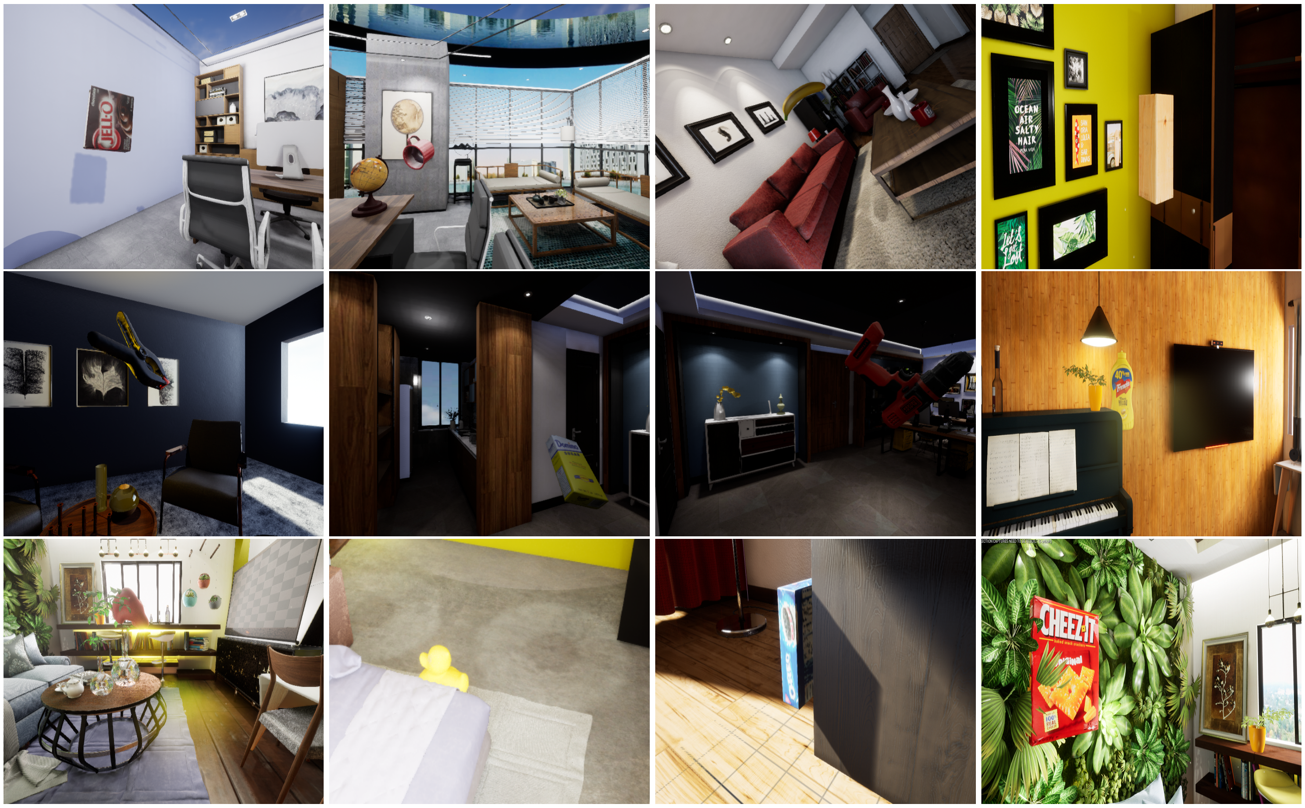


Figure 2. *Examples of objects in different scene types.*

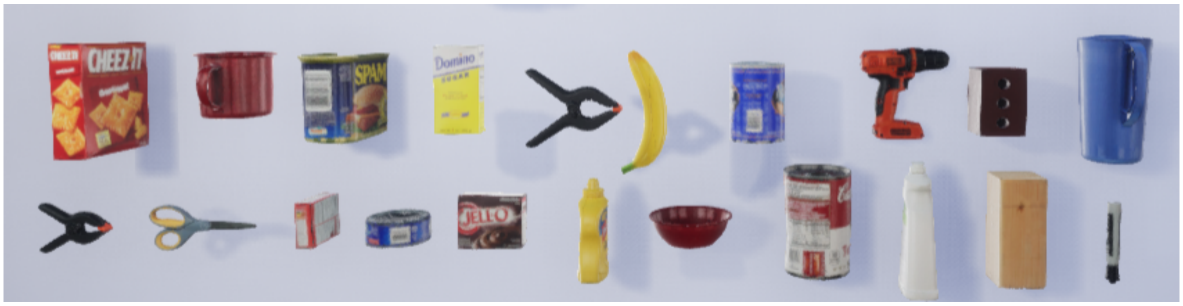


Figure 3. *3D object models in the YCB dataset.*

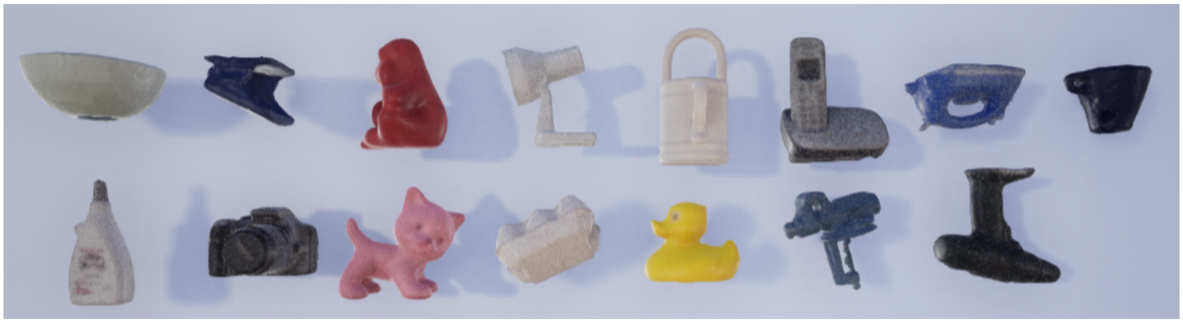


Figure 4. *3D object models in the LINEMOD dataset.*

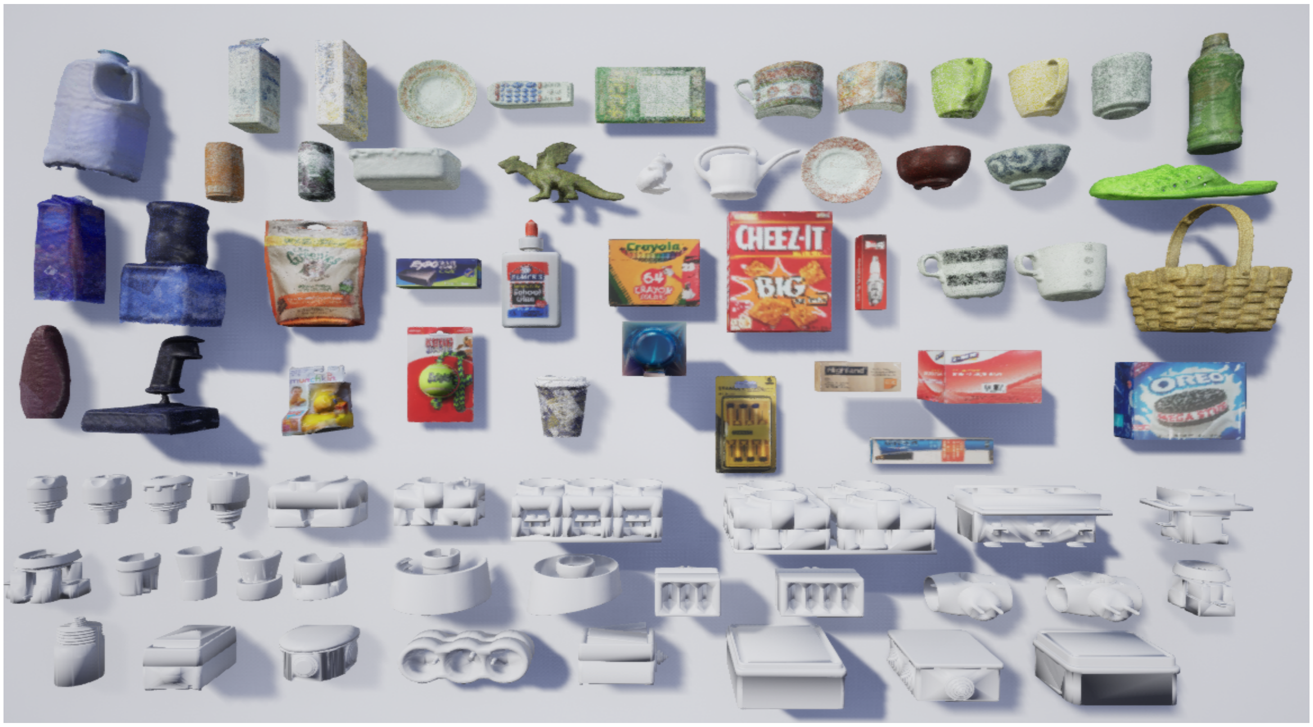


Figure 5. *3D object models in the TYO-L [6], TUD-L [6], IC-MI [17], RU-APC [16], and T-LESS [5] datasets.*

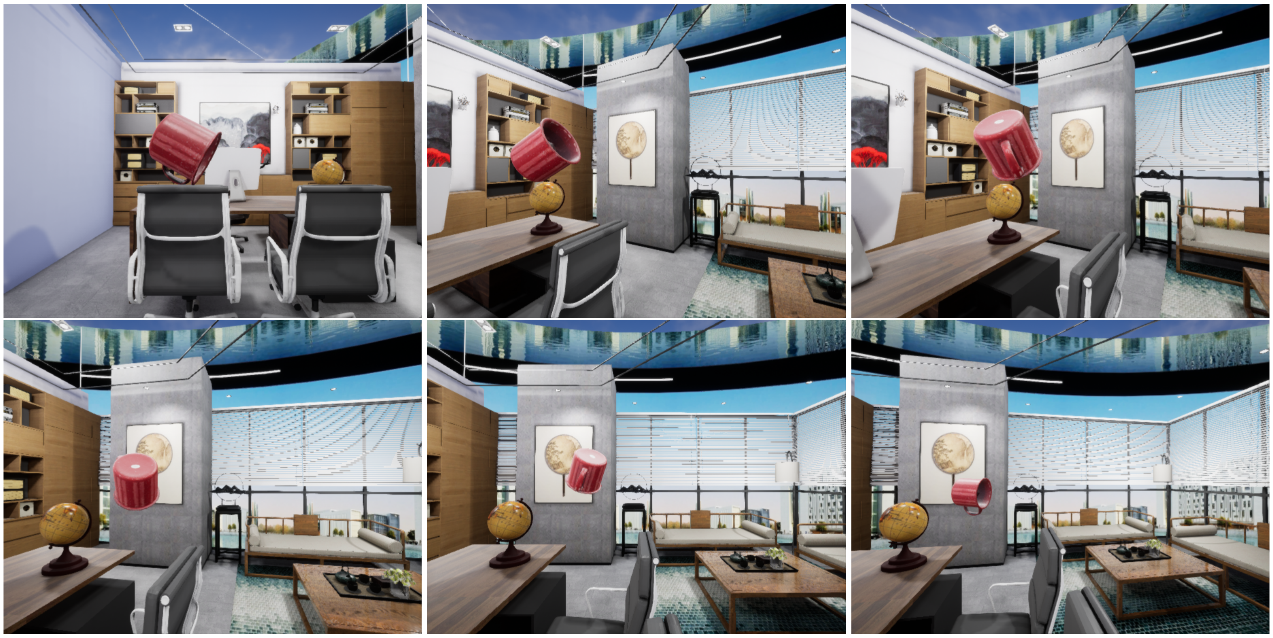


Figure 6. *Linear interpolation trajectory from candidate location points.*

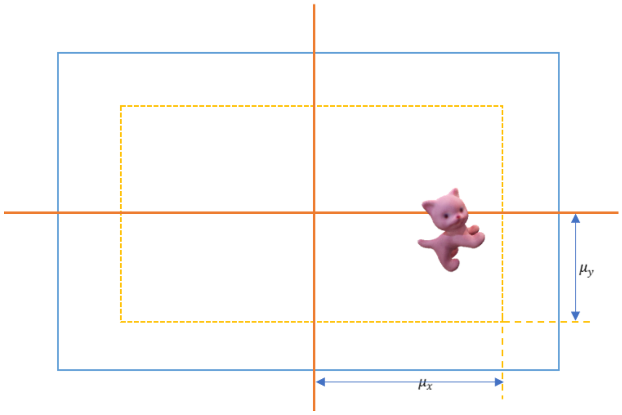


Figure 7. *Pixel coordinate constraint.*

[14] W. Qiu and A. Yuille, UnrealCV: Connecting Computer Vision to Unreal Engine, Proceedings of the European Conference on Computer Vision, pages 909–916, (2016)

[15] M. Rad and V. Lepetit, BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth, Proceedings of the IEEE International Conference on Computer Vision, pages 3848–3856, (2017)

[16] C. Rennie, R. Shome, K. E. Bekris, and A. F. De Souza, A Dataset for Improved RGBD-based Object Detection and Pose Estimation for Warehouse Pick-and-Place, IEEE Robotics and Automation Letters, 1(2), 1179–1185 (2016)

[17] A. Tejani, D. Tang, R. Kouskouridas, and T.-K. Kim, Latent-Class Hough Forests for 3D Object Detection and Pose Estimation, Proceedings of the European Conference on Computer Vision, pages 462–477 (2014)

[18] J. Tremblay, T. To, and S. Birchfield, Falling Things: A Synthetic Dataset for 3D Object Detection and Pose Estimation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 2038–2041 (2018)

[19] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox and S. Birchfield, Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects, Proceedings of the Conference on Robot Learning, (2018)

[20] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes, Proceedings of Robotics: Science and Systems, (2018)

Author Biography

Jianhang Chen is currently a Ph.D. candidate from Electrical and Computer Engineering, Purdue University. Before that, he joined NTU Robotics Lab, and received his M.S. degree in Mechanical Engineering from National Taiwan University (2017). He received his B.S. in Info. and Comp. Science from Beihang University (2014). His research focuses on computer vision and deep learning, with emphasis on printed image quality and registration.

Daniel Mas Montserrat graduated from Polytechnic University of Catalonia in 2015. He is currently a Ph.D. candidate at Purdue University under the supervision of Professor Edward J. Delp. He has worked as a research assistant applying machine learning in various projects funded by HP Labs and DARPA. His main areas of research are deep learning, signal processing, image processing and media forensics.

Dr. Qian Lin is a research scientist working on computer vision and deep learning research in HP Labs. She is also an Adjunct Full Professor of Electrical and Computer Engineering, Purdue University. She received her BS from Xian Jiaotong University in China, her MSEE from Purdue University, and her Ph.D. in EE from Stanford University. Dr. Lin is inventor/co-inventor for 44 issued patents. She was awarded Fellowship by the Society of IS&T in 2012, Outstanding Electrical Engineer by the School of ECE of Purdue University in 2013, and was promoted to the rank of HP Fellow in 2019.

Edward J. Delp was born in Cincinnati, Ohio. He is currently The Charles William Harrison Distinguished Professor of Electrical and Computer Engineering and Professor of Biomedical Engineering at Purdue University. In 2004 he received the Technical Achievement Award from the IEEE Signal Processing Society, in 2008 the Society Award from IEEE Signal Processing Society, and in 2017 the SPIE Technology Achievement Award. In 2015 he was named Electronic Imaging Scientist of the Year by IS&T and SPIE. Dr. Delp is a Life Fellow of the IEEE, a Fellow of the SPIE, and a Fellow of IS&T and a Fellow of the American Institute of Medical and Biological Engineering.

Jan P. Allebach received his B.S. from the University of Delaware in 1972, his M.S. from Princeton University in 1975 and his PhD from Princeton University in 1976. He is now the Hewlett-Packard Distinguished Professor of Electrical and Computer Engineering at Purdue University. Allebach is a Fellow of the IEEE, the National Academy of Inventors, the Society for IS&T, and SPIE. He was named Electronic Imaging Scientist of the Year by IS&T and SPIE, and was named Honorary Member of IS&T, the highest award that IS&T bestows. He has received the IEEE Daniel E. Noble Award, and is a member of the National Academy of Engineering. He recently served as an IEEE Signal Processing Society Distinguished Lecturer (2016-2017). His current research interests include image rendering, image quality, color imaging and color measurement, printer and sensor forensics, and digital publishing.

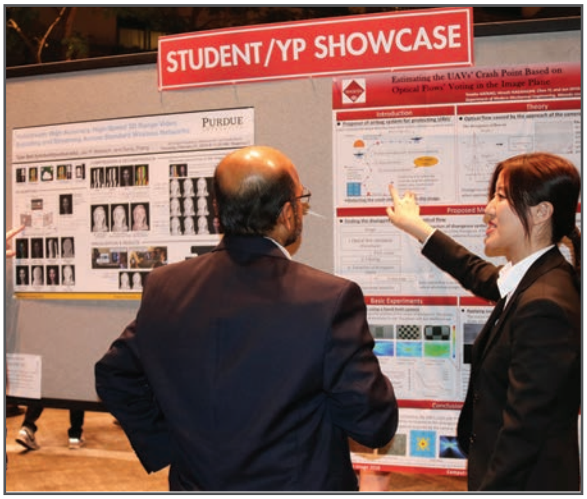
JOIN US AT THE NEXT EI!

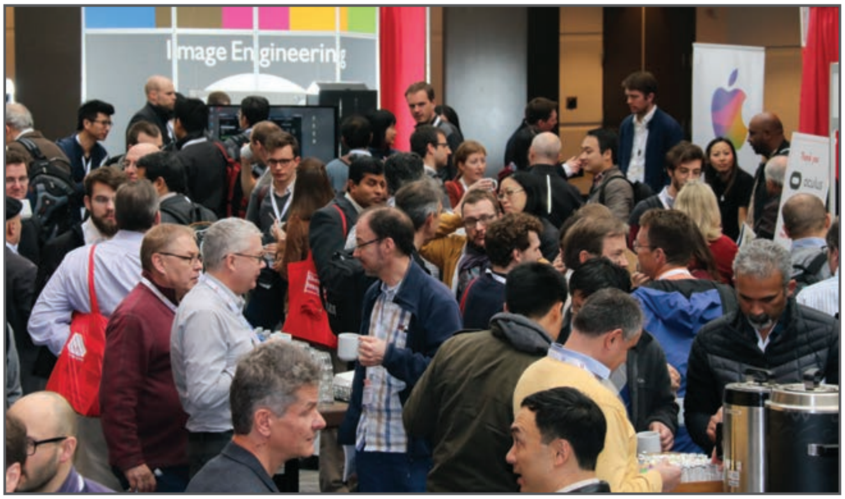
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





AR is a fundamental and transformative interface and platform.

20+ years from my initial working demo to today's AR displays

Optimistic that it won't take another 20 to reach ubiquitous consumer AR!

• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org