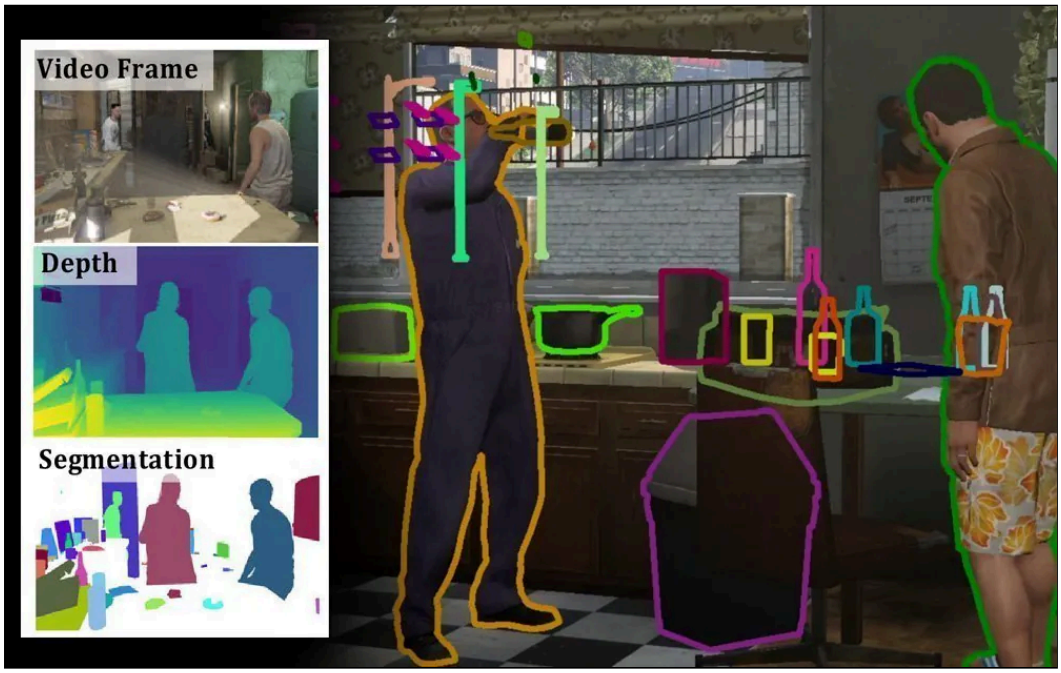
Synthetic Data: Bridging the Occlusion Gap With Grand Theft Auto

Originally published at https://www.unite.ai/synthetic-data-sail-vos-gta-v/

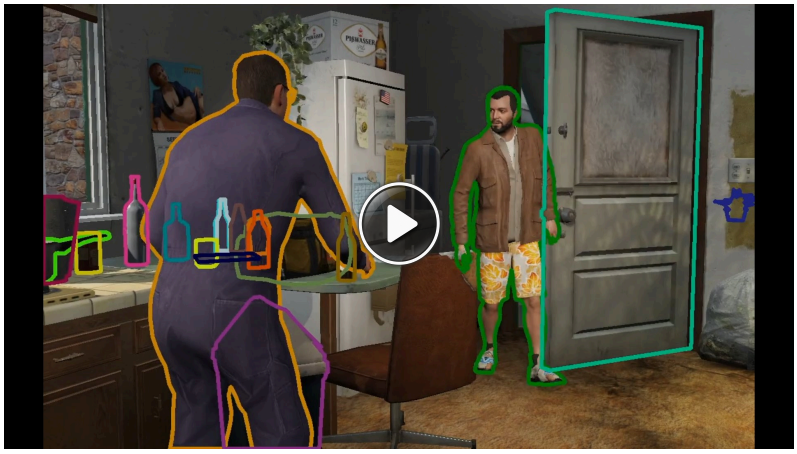


Researchers at the University of Illinois have created a new computer vision dataset that uses synthetic imagery generated by a Grand Theft Auto game engine to help solve one of the thorniest obstacles in semantic segmentation – recognizing objects that are only partly visible in source images and videos.

To this end, as described in the paper, the researchers have used the GTA-V video game engine to generate a synthetic dataset that not only features a record-breaking number of occlusion instances, but which features perfect semantic segmentation and labelling, and accounts for temporal information in a way that is not addressed by similar open source datasets.

Complete Scene Understanding

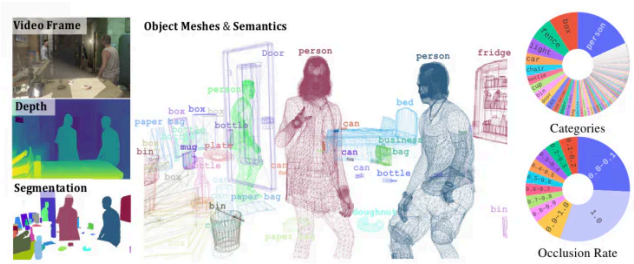
The video below, published as supporting material for the research, illustrates the advantages of a complete 3D understanding of a scene, in that obscured objects are known and exposed in the scene in all circumstances, enabling the evaluating system to learn to associate partial occluded views with the entire (labeled) object.



CLICK TO PLAY VIDEO ON SITE

Source: http://sailvos.web.illinois.edu/_site/index.html

The resulting dataset, called SAIL-VOS 3D, is claimed by the authors to be the first synthetic video mesh dataset with frame-by-frame annotation, instance-level segmentation, ground truth depth for scene views and 2D annotations delineated by bounding boxes.



Source (Click to enlarge)

The annotations of SAIL-VOS 3D include depth, instance-level modal and amodal segmentation, semantic labels and 3D meshes. The data includes 484 videos totaling 237,611 frames at 1280×800 resolution, including shot transitions.

Above, the original CGI frames; second row, instance-level segmentation; third row, amodal segmentation, which illustrates the depth of scene understanding and transparency available in the data. Source (Click to enlarge)

The set breaks down into 6,807 clips with an average of 34.6 frames each, and the data is annotated with 3,460,213 object instances originated from 3,576 mesh models in the GTA-V game engine. These are assigned to a total of 178 semantic categories.

Mesh Reconstruction And Automated Labeling

Since later dataset research is likely to occur on real world imagery, the meshes in SAIL-VOS 3D are generated by the machine learning framework, rather than derived from the GTA-V engine.

With a programmatic and essentially 'holographic' understanding of the entire scene representation, SAIL-VOS 3D imagery can synthesize representations of objects ordinarily hidden by occlusions, such as the far-facing arm of the character turning around here, in a way that would otherwise depend on many representative instances in real-world footage. (Click to enlarge) Source: https://arxiv.org/pdf/2105.08612.pdf

Since each object in the GTA-V world contains a unique ID, SAIL-VOS retrieves these from the rendering engine using the GTA-V script hook library. This solves the problem of reacquiring the subject if it should leave the field of view temporarily, as the labeling is persistent and dependable. There are 162 objects available in the environment, which the researchers mapped to a corresponding number of classes.

A Variety Of Scenes And Objects

Many of the objects in the GTA-V engine are common in nature, and therefore the SAIL-VOS inventory contains a fortunate 60% of the classes present in the Microsoft's frequently-used 2014 MS-COCO dataset.

The SAIL-VOS dataset includes a large variety of interior and exterior scenes under different weather conditions, with characters wearing varied clothing. (Click to enlarge)

Applicability

To ensure compatibility with the general run of research in this area, and to confirm that this synthetic approach can benefit non-synthetic projects, the researchers evaluated the dataset using the frame-based detection approach employed for MS-COCO and the 2012 PASCAL Visual Object Classes (VOC) Challenge, with average precision as the metric.

The researchers found that pre-training on the SAIL-VOS dataset improves the performance of Intersection over Union (IoU) by 19%, with a corresponding improvement in VideoMatch performance, from 55% to 74% on unseen data.

However, in cases of extreme occlusion, there were occasions when all of the older methods remained unable to identify an object or person, though the researchers forecast that this could be remedied in the future by examining adjacent frames to establish the reasoning for the amodal mask.

In the two right-hand images, traditional segmentation algorithms have failed to identify the female figure from the very limited portion of her head that's visible. Later innovations with optical flow evaluation may improve these results. (Click to enlarge)

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More