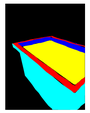
Bridging the Gap Between Real-World and Synthetic Domains in Semantic Segmentation

Sean Chen, Nico Gonnella, Nicholas Haisler, Haris Mehuljic, Khalid Mohammed

Supervised by Eric Manley and Alimoor Reza

Motivation



The diversity and volume of data in deep machine learning models are essential for enabling effective generalization to real-world scenarios. However, acquiring a sufficiently diverse dataset for computer vision applications—comprising tens of thousands of images—is often economically infeasible. In this research, we explored **domain randomization framework** [2] to mitigate this challenge by bridging the gap between large-scale synthetic data, which can be generated efficiently, and real-world data, which is often available only in limited quantities due to cost constraints. Our objective is to solve the semantic segmentation task, which involves predicting pixel-wise labels for an image. Specifically, we focus on segmenting various components of the grease bin, namely the ***inner-wall, outer-wall, top-surface, grease,*** and ***background***.

Domain Randomization for Semantic Segmentation Task

Synthetic data generated through computer graphics simulators is a viable approach and can often be produced at scale. However, discrepancies between simulated environments and the real world pose challenges for transferring knowledge from simulation to real-world applications. Building on the work of Tobin et al. [2], we **diversified the creation of our environments through randomization**, ensuring that data generated from these diverse environments can be utilized for training the deep neural network-based semantic segmentation model. We integrated a 3D model of a grease bin into various real-world 3D environment scans using **Blender** as a 3D simulation tool, *a widely used open-source 3D computer graphics software.* We randomize the following aspects of the domain for each sample during the synthetic data generation:

Category	Parameters	
Environments	3D scanned environment surrounding the bin.	
Camera Movement	Zoom	How close the camera is to the bin.
	Elevation	The vertical angle of the camera.
	Azimuth	The horizontal angle of the camera.
Textures	Bin Body	Rust & grunge levels on bin, as well as color.
	Grease	Specularity, chunkiness, and color of grease.
Grease Levels	The level of the grease inside the bin.	
Bin Positioning	The position of the bin within the environment.	

Real Data

Our real dataset consists of 179 images of various grease bins taken across the city. These images were collected from two sources: truck drivers responsible for grease collection and a student who worked on this project in previous years. Due to the dataset's limited size, it lacks the necessary variation to train a robust computer vision model. To develop a model capable of generalizing to diverse bin conditions, a larger and more varied dataset is required. Therefore, we chose to augment our dataset using computer-generated images.

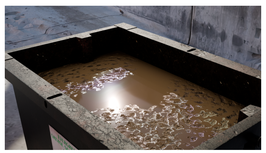
Generation of Synthetic Data with Semantic Segmentation Labels

To generate synthetic images resembling real data, we required three key components: a realistic bin model, a suitable environment for placement, and an efficient method for dataset generation. The bin model was created using CAD software, based on a dimensional drawing, to ensure accurate sizing. To obtain environment models, we conducted photogrammetric scans of various real-world locations around our campus. The datasets were then rendered using **Blender**. To evaluate the impact of domain randomization on model training, we identified several parameters that could be varied during dataset generation. A **Blender** plugin was developed to streamline the process of adjusting these parameters. Additionally, this plugin facilitated the rendering of both synthetic images and their corresponding segmentation labels.

3D Scans

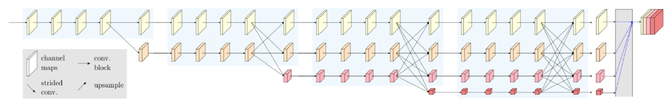
Our group selected 4 locations around Drake University where placing a bin would make sense. We photo-scanned these locations using the Polycam 3D app with the iPhone 14 Pro's LiDAR scanner to capture high-definition environment scans. These were exported into Blender, where the grease bin model was integrated to generate the training data.

Synthetic Dataset



We generated a dataset of approximately 4000 samples, which served as the basis for various experiments. The bin placement, environment, and bin/grease textures remained constant across all images in the dataset. In this initial experimental dataset, the only parameters varied were the distance between the camera and the bin, the horizontal and vertical angles of the camera relative to the bin, and the amount of grease in the bin.

Model Used

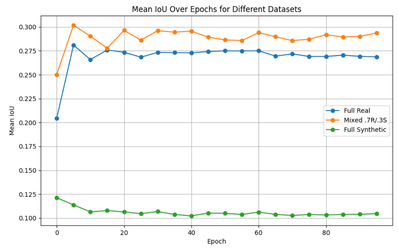


Model Used (Cont.)

HRNet's key strength as a model lies in its ability to preserve high-resolution information throughout the entire process. It achieves this by maintaining parallel high- and low-resolution convolutional streams, enabling the generation of representations that are both detailed and accurate. HRNet demonstrates its effectiveness across various applications, including human pose estimation, semantic segmentation, and object detection, making it a highly suitable pretrained model for our work [1].

Real and Synthetic Combination

To evaluate whether combining real and synthetic data enhances performance, we conducted a series of experiments. First, we trained HRNetV2 using only real images. Next, we trained the model using only synthetic images. Finally, we trained it with a combination of both, where 30% of the data was synthetic and 70% was real. Below we see the resulting graph, where the best model was trained using our mixed data method. We used mean intersection over union (mIoU) as our metric of performance for each model.



Future

Since we have demonstrated the effectiveness of our method, there are several directions in which we aim to extend this work. First, we seek to address the original problem of bin volume estimation using our improved model. This model can then be integrated into an application, enabling convenient estimation of grease volume, which has significant applications in renewable energy. Another approach involves the generation of synthetic data. Although we observed an improvement, the effect was marginal, likely due to the synthetic data not closely matching the real dataset. Enhancing the quality of synthetic data or introducing greater variation to improve our model remains a key area for future exploration.

References

[1] Wang, Jingdong and Sun, Ke and Cheng, et al. Deep high-resolution representation learning for visual recognition. IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI'18).

[2] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W. & Abbeel, P. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. IEEE/RSJ IROS'17.