TRoVE: Transforming Road Scene Datasets into Photorealistic Virtual Environments

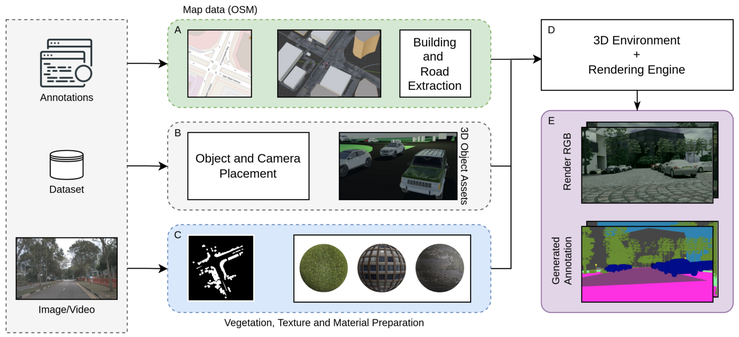
Shubham Dokania, Anbumani Subramanian, Manmohan Chandraker, C.V. Jawahar

Problem Statement

The aim is to generate synthetic data efficiently with high degree of photo-realism and multi-modal annotations. We propose **TRoVE** as a toolkit for generating synthetic data for real-world locations using existing dataset annotations.

Synthetic Data Generation

The figure below summarizes the process used in TRoVE for data generation process, which we also highlight in the following points:

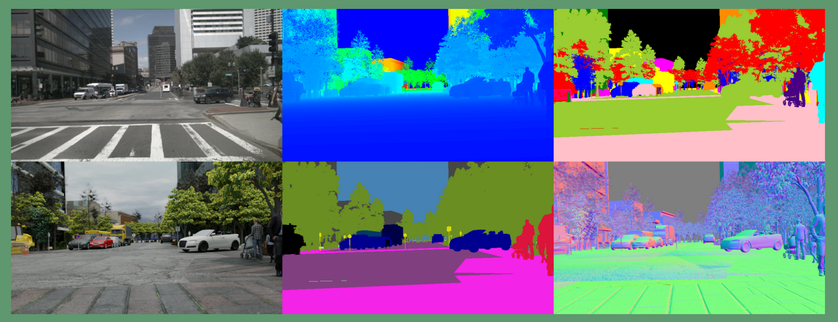


- **Virtual Environment**: We use OSM data and GPS location from annotated meta data to generate the Building and road layouts.
- **Camera and Objects**: For each available bounding box annotation, we select 3D object candidate fitting the box dimensions and place in the scene. Camera poses are simulated from meta-data if available otherwise randomly on vehicles.
- **Textures, Lighting & Background**: High-quality 4K PBR textures along with HDRI images for lighting are used for realistic scene layout. Vegeratation and Traffic poles/signs are spawned in the scene along the roads based on meta-data.

Scan for Toolkit Code Implementation, Dataset, Paper PDF, and 3D Assets used in the work.

https://github.com/shubham1810/trove_toolkit

TRoVE is a toolkit for effectively generating synthetic driving datasets from using existing datasets and annotations. Generating multi-modal datasets for various downstream tasks is one of the highlights in the proposed toolkit.



Highlighted samples from the data generation pipeline shows the multi-modal and diversity capabilities of the method for a similar scene layout with varying materials, lighting, colors and viewpoints.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD
CVIT
UNIVERSITY OF CALIFORNIA SAN DIEGO

ECCV TEL AVIV 2022

Experiments and Analysis

Training Method	Cityscapes		KITTI-STEP	
	mIoU	Accuracy	mIoU	Accuracy
R	70.25	98.88	59.81	98.39
S	30.63	95.50	27.64	92.96
S + R [F]	70.23	98.89	59.79	98.41
S + R [M]	70.82	98.93	**65.37**	**98.42**
S + C	37.03	95.84	46.14	95.92
S + C + R [F]	70.73	98.89	60.71	98.41
S + C + R [M]	**71.98**	**98.94**	63.93	98.38
P	61.44	98.65	56.44	98.28
S + P [F]	63.56	98.70	57.71	98.25
S + P [M]	**67.21**	**98.80**	61.72	**98.34**
S + C + P [F]	63.11	98.69	57.31	98.27
S + C + P [M]	65.75	98.78	**63.09**	98.33

The above table shows a brief report of the results from experiments performed on Cityscapes and KITTI-STEP datasets for semantic segmentation. The training method represents the training configuration where "S" is for synthetic, "R" for real, "P" is for partial-real data. "M" represents the training was done with mixed batch of synthetic and real, and "F" shows the real-world data was only used in fine-tuning. "C" represents usage of color-correction for color distribution alignment. Same results are visualized in the figure below for qualitative analysis.