

Generating Synthetic Data for Machine Learning

ICT has been generating synthetic data, which is artificially generated, and not based on real-world events, for machine learning for many years across projects including the following:

Human Behavior Understanding

Dr. Mohammad Soleymani's Intelligent Human Perception Lab has been developing multimodal sensing capabilities to analyze facial expressions, gaze and body posture. Together with his team, they have used synthetic data to improve facial expression analysis, action recognition and understanding of social behaviors in egocentric videos. Their work uses self-supervised learning (during pretraining) to learn rich representations that can be used across different downstream tasks related to expression and behavior understanding in videos.

Further reading:

- Contrastive Learning for Domain Transfer in Cross-Corpus Emotion Recognition (Yufeng Yin, Liupei Lu, Yao Xiao, Zhi Xu, Kaijie Cai, Haonan Jiang, Jonathan Gratch, Mohammad Soleymani)
- MagicPose: Realistic Human Poses and Facial Expressions Retargeting with Identity-aware Diffusion (Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, Mohammad Soleymani)

Geospatial Terrain: Semantic Terrain Points Labeling System (STPLS)

Dr. Andrew Feng's Geospatial Terrain Lab has developed a fully-automated workflow to segment 3D photogrammetric point clouds/meshes and extract object information, including individual tree locations and ground materials (Chen et al., 2019). The ultimate goal is to create realistic virtual environments and provide the necessary information for simulation. The generalizability of a previously proposed framework was tested using a database that was created under the U.S. Army's One World Terrain (OWT) project with a variety of landscapes (i.e., various buildings styles, types of vegetation, and urban density) and different data qualities (i.e., flight altitudes and overlap between images).

Although the database is considerably larger than existing databases, it remains unknown whether deep learning algorithms have truly achieved their full potential in terms of accuracy, as sizable data sets for training and validation are currently lacking. Obtaining large annotated 3D point cloud databases is time-consuming and labor-intensive not only from a data annotation perspective in which the data must be manually labeled by well-trained personnel but also from a raw data collection and processing perspective. Furthermore, it is generally difficult for segmentation models to differentiate objects, such as buildings and tree masses, and these types of scenarios do not always exist in the collected data set. Thus, the objective of this study is to investigate the possibility of using synthetic photogrammetric data to substitute for real-world data in training deep learning algorithms. The authors have investigated methods for generating synthetic UAV-based photogrammetric data to provide a sufficiently sized

database for training a deep learning algorithm with the ability to enlarge the data size for scenarios in which deep learning models have difficulties.

Further reading:

- [Generating Synthetic Photogrammetric Data for Training Deep Learning-based 3D Point Cloud Segmentation Models](#) (Meida Chen, Andrew Feng, Kyle McCullough, Pratusha Bhuvana Prasad, Ryan McAlinden, Lucio Soibelman)

Vision and Graphics Lab (VGL)

Much of Dr. Yajie Zhao's research for VGL is based on synthetic data, such as ML based virtual human projects (Deep Volumetric Video, Pifu, SiCloPe, PifuHD, ARCH, Portrait undistortion, PaGAN, HairNet). In the Face database project, where they used ground truth data, VGL also introduced synthetic data for data augmentation to make the networks robust.

Synthetic data is efficient for network prototype testing because it's clean and fully controllable, however, it introduces a domain gap if directly applied to the model trained with synthetic to real data. VGL's strategy is to enhance the sim2real accuracy with the introduction of depth or 3D information - compared to 2D, the domain gap of 3D between synthetic and real data is very small.

Synthetic data is significant for many ML tasks (especially supervised learning) because: (1) manual annotation of real data is expensive; (2) due to the limitations of the capture equipment, some specific training data is impossible or difficult to obtain (e.g. long-range depth, multi-view human performance data). In contrast, synthetic data with computer-generated annotations is much easier to generate. Despite this, the domain mismatch between the real data and the synthetic data may significantly decrease the learning models' performance because of the low quality and limited diversity of synthetic data.

The impact of domain gaps on ML depends on data quality and specific tasks, and it can usually be minimized in two ways. First, for some tasks on the face and body like those in our lab, the problem can be solved by improving the quality of synthetic data with delicate devices or advanced rendering algorithms. For example, the face models generated using light stage data and rendered with high-resolution textures can achieve comparable fidelity to real data, resulting in negligible domain gaps. Second, for general tasks such as scene understanding, it seems impossible to generate perfect synthetic data. The most recent solution is either to make the distribution of the composition closer to the real data by a domain transfer method or extract the domain invariant features of the synthetic data using domain adaptation techniques.

Further reading:

- [Sparse Multi-View Performance Capture ECCV](#)
 - [SiCloPe: Silhouette-Based Clothed People](#)
 - [PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization](#)
 - [Learning Perspective Undistortion of Portraits](#)
 - [Learning Formation of Physically-Based Face Attributes](#)
 - [HairNet: Single-View Hair Reconstruction using Convolutional Neural Networks](#)
 - [Intuitive, Interactive Beard and Hair Synthesis with Generative Models](#)
-

Master Scenario Events List Injection (MSELI)

In MSELI, Dr. Andrew Gordon and Dr. Andrew Feng used synthetic training data of squad-level group behaviors (generated in RIDE) to recognize these behaviors when they are executed by teams of human players (in RIDE, again). This project used a machine-learning (ML) framework for recognizing tactical events in virtual training environments. Approach, unit movements, surrounding environments, and other atomic events are rasterized as a 2D image, allowing researchers to solve the action detection problem as image classification and video temporal segmentation tasks. In order to bootstrap ML models, synthetic training data to procedurally generate a large amount of annotated data was utilized, demonstrating the effectiveness of this framework in the context of a virtual military training prototype for detecting troop formations.

Strategy Optimization in Deep Multi-agent Reinforcement Learning for Military Training Simulations (SOMATS)

Within Dr. Volkan Ustun's Human-inspired Adaptive Teaming Systems the SOMATS project utilizes synthetic training data generated by leveraging captured decision flows of SMEs for a squad level scout mission scenario. Such decision flows are then utilized in reinforcement learning experiments to guide the predefined behaviors in the simulation environments. Lack of annotated training data for military applications make such an approach necessary for AI models. Furthermore, reinforcement learning techniques rely on the experiences generated in simulation environments. Representations of the environment, synthetic agent behaviors, outcomes of the actions taken in the simulation environment determine the generated synthetic experiences, which in turn generates the signal to update and better the behavior policies of the synthetic characters.

Further reading:

- [Multi-agent Reinforcement Learning with a Scout Mission Scenario](#) (Volkan Ustun, Rajay Kumar, Lixing Liu, Nicholas Patitsas)