# Title & Team

**Overcome data scarcity for transformer-based vision models.**

Adrian Brunner, [adrian.brunner@stud.hslu.ch](mailto:adrian.brunner@stud.hslu.ch), data & modelling

Chris Imholz, [chris.imholz@stud.hslu.ch](mailto:chris.imholz@stud.hslu.ch), data & modelling

Tim Zoller, [tim.zoller@stud.hslu.ch](mailto:tim.zoller@stud.hslu.ch), modelling & evaluation

# Introduction / Problem Statement

Data scarcity is one of the major bottlenecks in modern computer vision. Transformer-based vision models, such as DETR, have shown remarkable results in object detection, but they typically require large amounts of annotated data to reach their full potential. Collecting and labeling such datasets in the real world is expensive, time-consuming, and error-prone, especially for safety-critical domains like autonomous driving. This challenge motivates the exploration of synthetic datasets, which can be generated automatically at scale.

Grand Theft Auto V (GTA V), with its highly detailed, photo-realistic world, has proven to be a valuable tool for generating synthetic data for vision research. The PreSIL dataset leverages GTA V to provide precisely annotated synthetic images and LiDAR data, enabling large-scale experiments without manual labeling.

**Research Question / Hypothesis**

RQ1. Can synthetic data generated from GTA V (PreSIL dataset) improve the performance of transformer-based vision models for object detection?

**Out of Scope**:

* Pose estimation, semantic/instance segmentation, multi-modal fusion and real-time deployment considerations.
* Expanding object classes beyond cars and people.

# Related Work

Synthetic datasets have been leveraged to address data scarcity in computer vision by providing large-scale, automatically annotated data. PreSIL created by Hurl et al. (2019) uses the GTA V engine to generate over 50 k frames with high-resolution images, precise 2D and 3D annotations for vehicles and pedestrians, and automatic LiDAR and depth—demonstrating notable improvements (≈ 5% AP) on the KITTI 3D object detection benchmark when combined with real data. Similarly, ViPeD (Ciampi et al., 2020) employs synthetic pedestrian images from GTA V and explores methods to reduce the synthetic-to-real domain gap for pedestrian detection.

Most prior work focuses on CNN-based detectors or applies synthetic data to pretraining stages, few studies evaluate transformer-based detectors like DETR. Our project bridges this gap by conducting a fine-tuning–only study of transformer-based detection models on synthetic (PreSIL) and real data.

# Data

**PreSIL (GTA V) (Access currently requested)**

Synthetic dataset in 2D RGB (LiDAR available but unused here); approximately 50k instances; precise bounding boxes for vehicles and pedestrians; automatically generated labels; research license.

**Label Quality**: PreSIL provides highly precise automatic annotations; KITTI contains carefully curated manual annotations (benchmark quality).

**Cityscapes (Real Data) (**[***gtCoarse.zip***](https://www.cityscapes-dataset.com/file-handling/?packageID=2)**)**  
Real-world dataset in 2D RGB street scenes; recorded across 50 cities in Europe (focus on urban driving); 5k finely annotated images (with 30 classes) and ~20k coarsely annotated images; provides pixel-level semantic, instance, and panoptic segmentation; research license.

**Label Quality:** Cityscapes fine annotations are of very high manual quality (benchmark standard), while the coarse labels cover larger scale but are less detailed.

# Method / Proposed Approach

A diagram of a model

AI-generated content may be incorrect.We aim to investigate whether synthetic data can improve transformer-based object detection models under data scarcity. Instead of pretraining from scratch, we will use transfer learning with fine-tuning on existing state-of-the-art architectures. Specifically, we will experiment with RT-DETR (a real-time variant of DETR) and YOLOv10, both of which provide competitive detection performance with reasonable computational costs.

Our study consists of three training regimes:

* R1: Fine-tuning on synthetic PreSIL data only.
* R2: Fine-tuning on real KITTI data only (restricted to *car* and *person* classes).
* R3: Fine-tuning on a synthetic-heavy mixture (≈ 90% synthetic, 10% real). This unbalanced setup reflects a realistic scenario where collecting a large real dataset is expensive, and synthetic data provides the bulk of training examples.

For efficiency, we will initialize models with COCO-pretrained weights. Fine-tuning will be combined with data augmentation techniques such as random cropping, color jittering, and weather effects to reduce overfitting and mitigate domain shift. Training is designed to be computationally feasible, with expected runtimes of a few hours per regime on our or Google Colabs GPUs.

Evaluation will be based on COCO-style metrics (mAP, AP50, AP75) as well as class-wise AP for *car* and *person*.

# References

Ciampi, L., Messina, N., Falchi, F., Gennaro, C., & Amato, G. (2020). Virtual to Real adaptation of Pedestrian Detectors. Sensors, 20(18), 5250. https://doi.org/10.3390/s20185250

Hurl, B., Czarnecki, K., & Waslander, S. (2019, Mai 1). Precise Synthetic Image and LiDAR (PreSIL) Dataset for Autonomous Vehicle Perception. arXiv.Org. https://arxiv.org/abs/1905.00160v2

**Freigabe**

|  |  |
| --- | --- |
| Place, Date |  |
| Lecturer |  |