1. INTRODUCTION

In the automotive setting, safety is a mandatory requirement. Current literature approaches are still end-to-end, which tend to obscure the learned knowledge. Modelling a method with explainable operations, it is certainly an advantage in terms of acting as similar as possible to the human way of thinking.

In this work, we propose a two stages framework able to generate realistic visual futures for urban scenes with vehicles as main actors:
1. From raw RGB frames we extract interpretable information including bounding boxes and trajectory estimation;
2. We produce visual intermediate inputs and feed them into a deep neural network to generate the final visual appearance of the vehicle in the future.

This approach is more suitable for the human-vehicle interaction setting with intermediate representation that a human can understand and interact with naturally. Moreover, the input resolution does not represent a limit because, since only individual vehicle images are processed, their resolution is typically lower than the full frame.

To sum up, our pipeline:
- exploits intermediate high-level interpretable information to produce deterministic visual future grounded on those constraints;
- is not limited to unimodal input, but can generate alternative futures depending on the given input;
- outperforms end-to-end methods both qualitatively and quantitatively.

INTERPRETABLE INFORMATION EXTRATION

- Vehicle detection: an SSD network outputs detected bounding boxes
- Trajectory prediction: a graph-based network, TrackletNet, performs a tracking-by-detection algorithm
- Keypoints detection: a Stacked Hourglass network outputs 12 semantic keypoints (e.g. wheels, lights, window corners)
- 3D model classification: a VGG19 network classifies the vehicle into 10 possible 3D synthetic models
- 3D pose estimation: a Levensberg-Marquardt optimization algorithm computes the vehicle 6DoF pose
- Trajectory rototation: the 3D lifted predicted trajectory (pixel to GPS/meters) is used to compute consecutive rotation transformations of the vehicle from its starting position to its future position (+1s)

NOVEL VIEW COMPLETION

- An image completion network takes as input a rototranslated 3D synthetic model and the initial vehicle appearance in the first frame and outputs a realistic textured synthetic model from the new viewpoint

GROUND TRUTH

\[ t = 0 \]

PREDICTION

\[ t + 0.2s \]
\[ t + 0.6s \]
\[ t + 1s \]

2. PROPOSED METHOD

3. DATASETS

REFERENCES


4. RESULTS

We compare our method with end-to-end image-to-image translation and recurrent approaches. Our framework outperforms those end-to-end baselines according to several appearance metrics focused on the vehicle bounding box area.

PASCAL3D+

It is a collection of images from 12 different object classes with annotations of 2D keypoints, 3D synthetic model class (10 for the vehicle class) and 3D pose.

CARFUSION

It contains videos of traffic intersections taken by people on a sidewalk. Each frame has annotated bounding boxes and 2D keypoints for each vehicle.

CITYFLOW

It represents a challenging traffic surveillance dataset with annotations of vehicles detection, tracking and re-identification.

CODE

Project code available on GitHub: