Future Urban Scene Generation Through Vehicle Synthesis

Alessandro Simoni, Luca Bergamini, Andrea Palazzi, Simone Calderara, Rita Cucchiara

{alessandro.simon, luca.bergamini24, andrea.palazzi, simone.calderara, rita.cucchiara}@unimore.it

University of Modena and Reggio Emilia, Italy
In the literature traffic monitoring and autonomous driving problems are usually addressed with end-to-end methods.

Since safety is a mandatory requirement, the method interpretability should be as similar as possible to the human way of thinking\textsuperscript{1,2}

In this work, we present a novel two stages framework reproducing deterministic visual future for videos taken by traffic surveillance cameras.

We show how our method can output “alternative futures” depending on the given inputs and how it outperforms end-to-end image-to-image translation and recurrent approaches.

In the automotive setting RGB cameras are certainly an enabling technology for scene understanding tasks.

Most of the literature approaches rely also on LiDAR/radar or depth sensors which capture precise 3D information of the scene.

Our challenging goal is to extract information about vehicles from monocular RGB data only and use them to generate a 3D synthetic representation reprojected into the scene.

In addition to this, we also consider the temporal trajectory within 1s in the future of each vehicle rotating and translating them from their start position in the frame.

The given output will be a video representing the same scene in which each vehicle is replaced with its synthetic textured model.
• A key feature of our approach relies on the correspondence between 2D predicted keypoints and 3D annotated keypoints on 3D synthetic vehicle models.

• This information enables the computation of an object viewpoint with respect to a camera point of view, the well-known perspective-n-point problem.
Overview of the proposed method

**Proposed method**

1. **Novel View Completion**
   - Target appearance
   - Source appearance
   - Static background
   - Image Completion Network

2. **Interpretable Information Extraction**
   - Trajectory Prediction
   - Keypoints Detection
   - 3D Model Classification
   - 3D Pose Estimation

3. **Vehicles Detection**
4. **3D Pose Estimation**
5. **3D Model Classification**
6. **Keypoints Detection**
7. **Trajectory Prediction**

**Overview**

- At time $t$
- Target 3D Normals

- At time $t + 0.2s$
Proposed method

**INTERPRETABLE INFORMATION EXTRACTION**

- **Input:**
  set of N frames from RGB camera device

- **Vehicle detection:**
  an SSD\(^1\) architecture outputs vehicle bounding boxes

- **Trajectory prediction:**
  a graph-based network, TrackletNet\(^2\), performs a tracking-by-detection algorithm

Proposed method

INTERPRETABLE INFORMATION EXTRACTION

- **Input:**
  vehicle image cropped from its bounding box

- **3D model classification:**
  a VGG19$^1$ network outputs the 3D vehicle model class

- **2D keypoints localization:**
  a Stacked Hourglass$^2$ architecture outputs 12 semantic keypoints (e.g. wheels, lights, frontal and back window corners)

---

Proposed method

INTERPRETABLE INFORMATION EXTRACTION

- **Input:** annotated 3D keypoints, predicted 2D keypoints and trajectory

- **3D pose estimation:** a Levenberg-Marquardt\(^1\) pose optimization iterative algorithm outputs the initial 6DoF vehicle pose

- **Trajectory rototranslation:**
  
  3D lifted trajectory (pixel to GPS/meters) is applied as consecutive transformations

---

Proposed method

NOVEL VIEW COMPLETION

• **Input:**
  trajectory rototranslation, 3D vehicle model, cropped vehicle image

• **Novel view completion:**
  an image completion network\(^1,^2\) exploits appearance information from initial vehicle image and 3D normals of the rototranslated model and outputs vehicle appearance from the new viewpoint

Datasets

Pascal3D+\(^1\)
- Collection of images from 12 different object classes
- Annotations of 2D keypoints, 3D model class, 3D pose
- 10 possible 3D synthetic vehicle models

CarFusion\(^2\)
- Videos of street intersections taken by people on a sidewalk
- Annotations of bounding boxes and 2D keypoints for each vehicle

CityFlow\(^3\)
- Videos of street intersections taken from traffic surveillance cameras
- Annotations of detection, tracking and re-identification information

---

• Image-to-image translation and recurrent baseline networks

Qualitative results

• Our approach

We compare our results on the CityFlow dataset evaluating the difference in the appearance of the cropped area of each vehicle.

The proposed approach outperforms both image-to-image translation and recurrent baseline networks where the results tend to be blurry or faded.

Our method maintains good performance in the long run throughout the entire temporal window in analysis (i.e. 1s in the future).
Conclusions

• We propose a novel framework for predicting visual future appearance of an urban scene
• As an alternative to end-to-end methods, we include human interpretable information and each actor in the scene is modelled independently
• We show how our method outperforms end-to-end approaches both qualitatively and quantitatively

Open issues:
• Improving 3D model classification accuracy avoiding class swapping
• Introducing some road constraints improving potential wrong initial poses

Code is available online: https://github.com/alexj94/future_urban_scene_generation
Thank you for your attention

Future Urban Scene Generation Through Vehicle Synthesis

Alessandro Simoni, Luca Bergamini, Andrea Palazzi, Simone Calderara, Rita Cucchiara

{alessandro.simoni, luca.bergamini24, andrea.palazzi, simone.calderara, rita.cucchiara}@unimore.it

University of Modena and Reggio Emilia, Italy