Future Urban Scene Generation Through Vehicles Synthesis

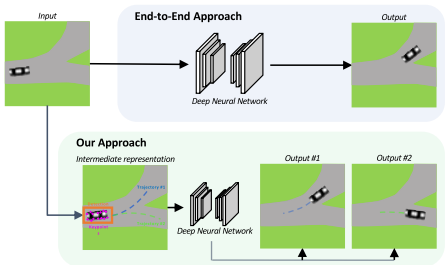
Alessandro Simoni, Luca Bergamini, Andrea Palazzi, Simone Calderara, Rita Cucchiara

Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia

ICPR 2020 — 25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION — Milan, Italy 10 | 15 January 2021

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.

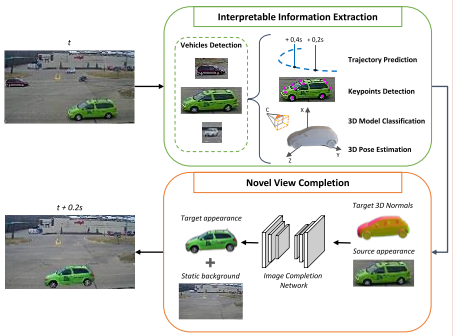
To sum up, our pipeline:

- 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.

REFERENCES

[1] T.-C. Wang et al., High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018.
[2] P. Esser et al., A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018.
[3] A. Palazzi et al., Warp and learn: Novel views generation for vehicles and other objects. In *TPAMI*, 2020.
[4] W. Lotter et al., Deep predictive coding net-works for video prediction and unsupervised learning. In *arXiv preprint arXiv:1605.08104*, 2016.

2. PROPOSED METHOD



INTERPRETABLE INFORMATION EXTRACTION

- **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 Levenberg-Marquardt optimization algorithm computes the vehicle 6DoF pose
- **Trajectory rototranslation:** the 3D lifted predicted trajectory (pixel to GPS/meters) is used to compute consecutive rototranslation 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*

3. DATASETS

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.

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.

TABLE I
COMPARISON ON THE TEST SET USING MEAN SQUARED ERROR (MSE). EACH COLUMN REFERS TO A FUTURE DISPLACE. LOWER IS BETTER.

Method	*+0.2s*	*+0.4s*	*+0.6s*	*+0.8s*	*+1.0s*
Pix2PixHD [1]	4854	4919	4950	4966	5007
Pix2Pix-3D	3579	3802	4026	4198	4424
PredNet [4]	**2037**	**2499**	2765	2877	2959
Our(VUnet [2])	2705	2692	**2759**	**2755**	**2870**
Our(Warp&Learn [3])	2996	2987	3058	3055	3153

TABLE II
COMPARISON ON THE TEST SET USING STRUCTURAL SIMILARITY INDEX (SSIM). EACH COLUMN REFERS TO A FUTURE DISPLACE. HIGHER IS BETTER.

Method	*+0.2s*	*+0.4s*	*+0.6s*	*+0.8s*	*+1.0s*
Pix2PixHD [1]	0.18	0.17	0.17	0.18	0.18
Pix2Pix-3D	0.24	0.23	0.22	0.21	0.20
PredNet [4]	0.40	0.37	0.35	0.36	0.36
Our(VUnet [2])	**0.50**	**0.50**	**0.49**	**0.50**	**0.50**
Our(Warp&Learn [3])	**0.50**	**0.50**	**0.49**	0.49	0.49

TABLE III
COMPARISON ON THE TEST SET USING INCEPTION SCORE (IS). EACH COLUMN REFERS TO A FUTURE DISPLACE. HIGHER IS BETTER

Method	*+0.2s*	*+0.4s*	*+0.6s*	*+0.8s*	*+1.0s*
Pix2PixHD [1]	2.27	2.29	2.32	2.32	2.24
Pix2Pix-3D	2.72	2.67	2.56	2.37	2.51
PredNet [4]	3.16	3.16	3.16	3.00	3.00
Our(VUnet [2])	**3.17**	**3.22**	**3.21**	**3.10**	**3.32**
Our(Warp&Learn [3])	2.93	2.78	2.87	3.02	2.91

TABLE IV
COMPARISON ON THE TEST SET USING FRECHET INCEPTION DISTANCE (FID). EACH COLUMN REFERS TO A FUTURE DISPLACE. LOWER IS BETTER

Method	*+0.2s*	*+0.4s*	*+0.6s*	*+0.8s*	*+1.0s*
Pix2PixHD [1]	274.2	268.6	265.3	262.3	259.4
Pix2Pix-3D	240.6	241.2	248.6	245.9	249.5
PredNet [4]	197.1	197.2	196.4	193.4	196.3
Our(VUnet [2])	192.8	187.3	182.0	178.3	177.49
Our(Warp&Learn [3])	**90.4**	**90.22**	**91.2**	**92.6**	**94.1**

CODE

Project code available on GitHub: