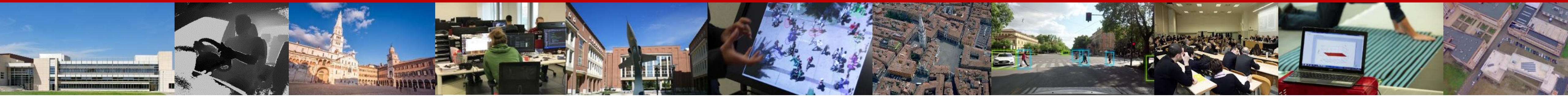
ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

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

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

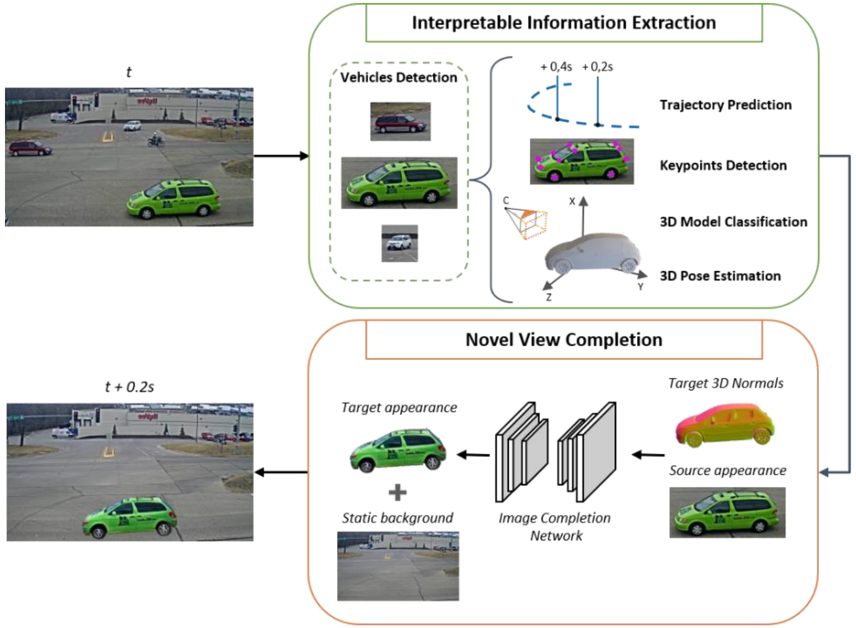
AImage Lab

Centro Interdipartimentale di Ricerca
Softech: ICT per le Imprese

Our Proposal

- 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**[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



1. *M. Bansal et al., Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. In arXiv:1812.03079, 2018.*
2. *J. Hong et al., Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions. In CVPR, 2019.*

Our Goal

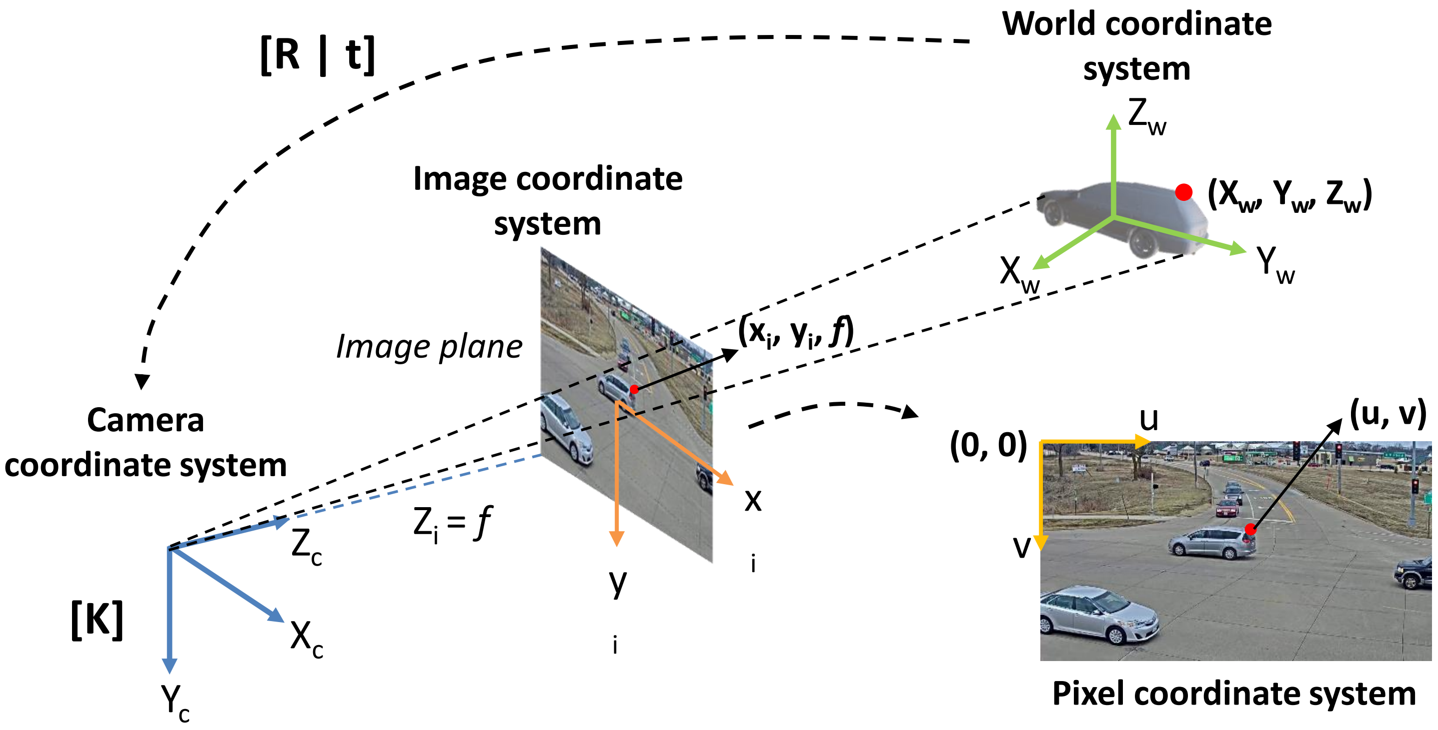
- In the automotive setting **RGB cameras** are certainly an enabling technology for **scene understanding tasks**
- Most of the literature approaches rely also on **LiDAR/radar or depth sensors** which capture precise 3D information of the scene

- 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**

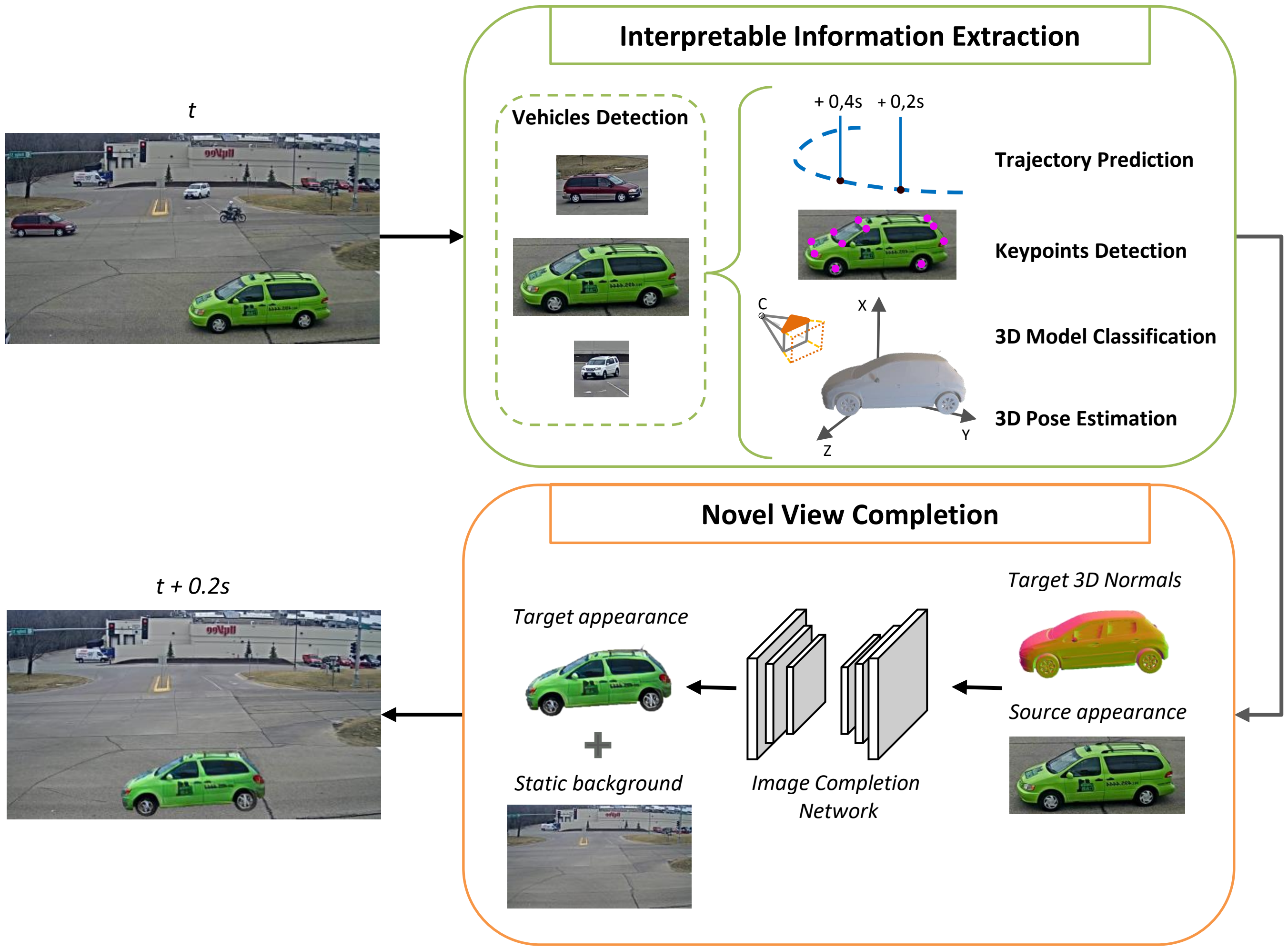
Keypoints and viewpoint

- 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



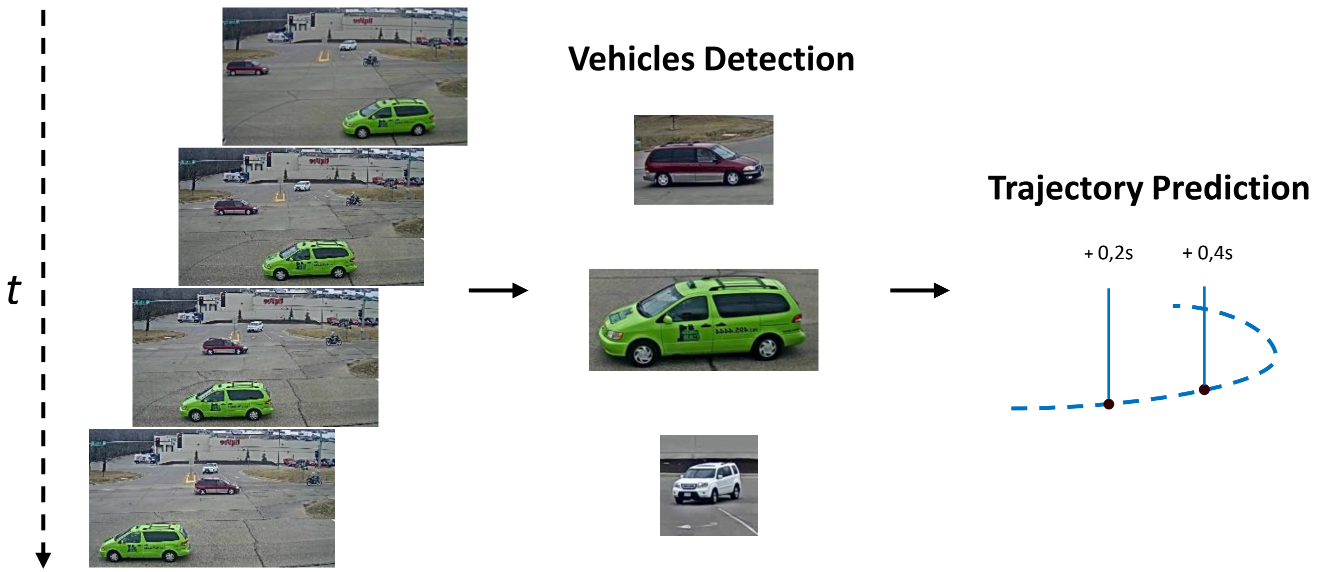
Proposed method

Overview of the proposed method



Proposed method

INTERPRETABLE INFORMATION EXTRACTION



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

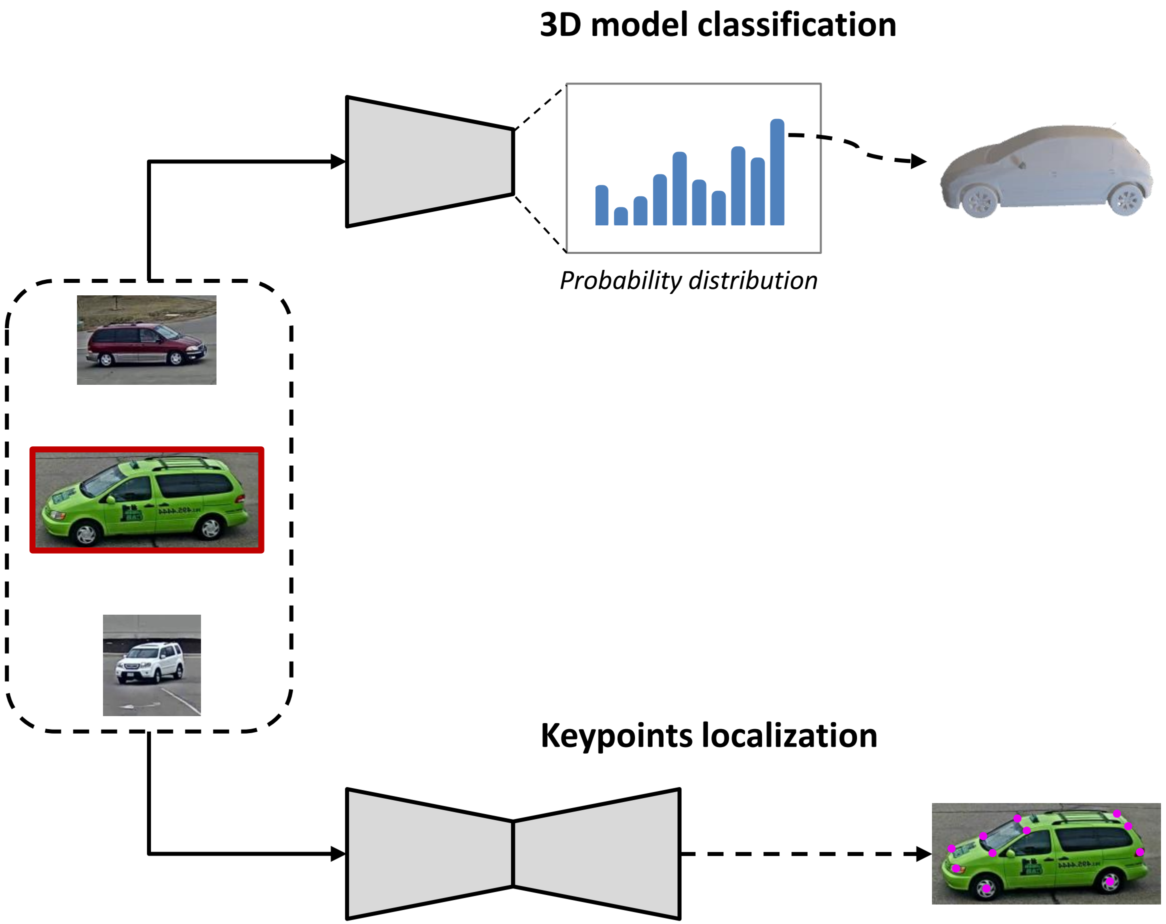
- **Vehicle detection:**
  an SSD[1] architecture outputs vehicle bounding boxes

- **Trajectory prediction:**
  a graph-based network, TrackletNet[2], performs a tracking-by-detection algorithm

1. *W. Liu et al., Ssd: Single shot multibox detector. In ECCV, 2016.*
2. *G. Wang et al., Exploit the connectivity: Multi-object tracking with trackletnet. In ACMMM, 2019.*

Proposed method

INTERPRETABLE INFORMATION EXTRACTION



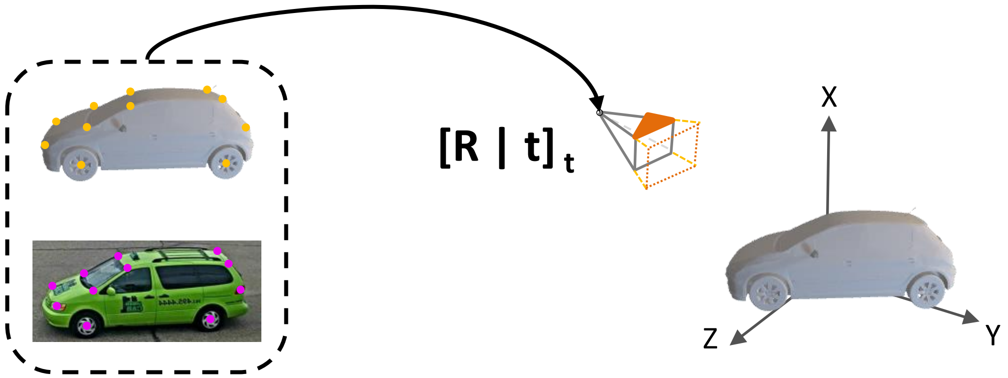
- **Input:**
  vehicle image cropped from its bounding box
- **3D model classification:**
  a VGG19[1] network outputs the 3D vehicle model class
- **2D keypoints localization:**
  a Stacked Hourglass[2] architecture outputs 12 semantic keypoints (e.g. wheels, lights, frontal and back window corners)

1. *K. Simonyan et al., Very deep convolutional networks for large-scale image recognition. In arXiv preprint arXiv:1409.1556, 2014.*
2. *A. Newell et al., Stacked hourglass networks for human pose estimation. In ECCV, 2016.*

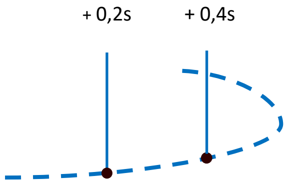
Proposed method

INTERPRETABLE INFORMATION EXTRACTION

3D pose estimation



Trajectory rototranslation



$$\{ [R \mid t]_{t}, [R \mid t]_{t+0.2s}, [R \mid t]_{t+0.4s} \}$$

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

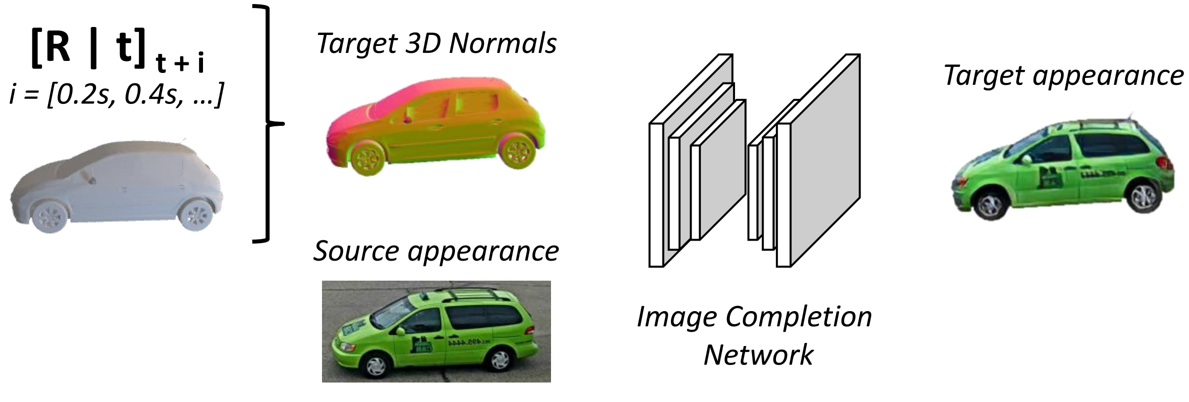
- **3D pose estimation:**
  a Levenberg-Marquardt[1] pose optimization iterative algorithm outputs the initial 6DoF vehicle pose

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

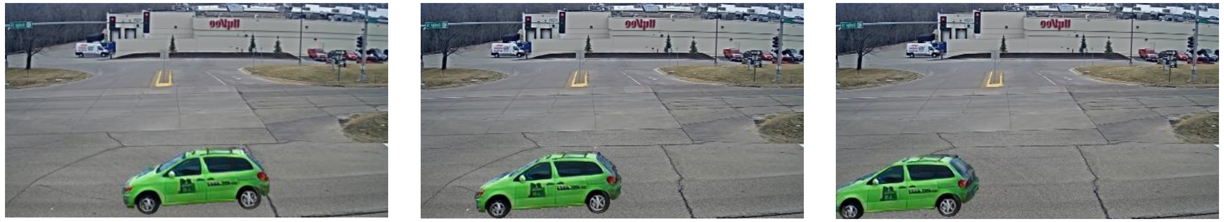
1. *M. Darcis et al., Poselab: A Levenberg-Marquardt based prototyping environment for camera pose estimation. In CISP-BMEI, 2018.*

Proposed method

NOVEL VIEW COMPLETION



Results at time *t+i*



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

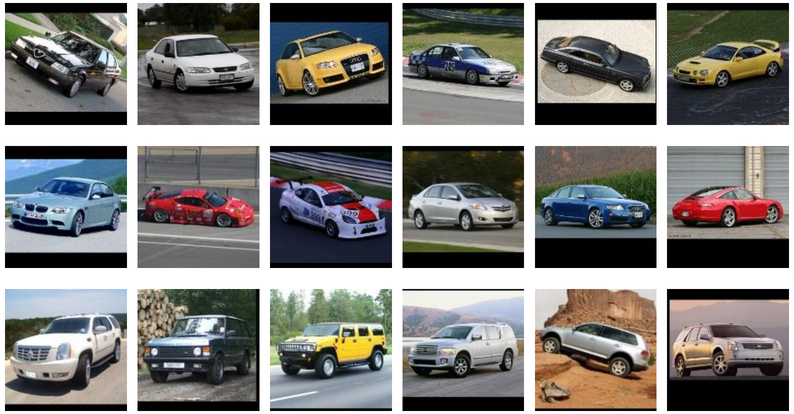
- **Novel view completion:**
  an image completion network[1,2] exploits appearance information from initial vehicle image and 3D normals of the rototranslated model and outputs vehicle appearance from the new viewpoint

1. P. Esser et al., *A variational u-net for conditional appearance and shape generation. In CVPR, 2018.*
2. A. Palazzi et al., *Warp and learn: Novel views generation for vehicles and other objects. In TPAMI, 2020.*

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

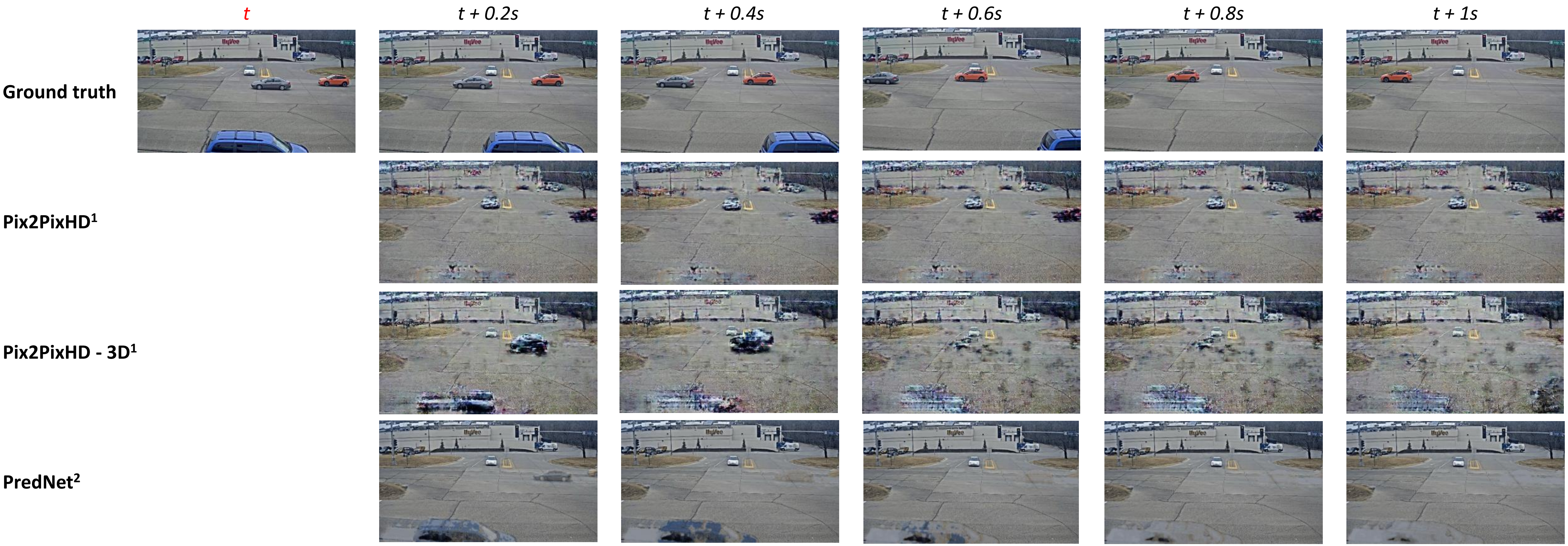
CityFlow[3]

- Videos of street intersections taken from traffic surveillance cameras
- Annotations of detection, tracking and re-identification information

1. *Y. Xiang et al., Beyond pascal: A benchmark for 3d object detection in the wild. In WACV, 2014.*
2. *N. Dinesh Reddy et al., Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicles. In CVPR, 2018.*
3. *Z. Tang et al., Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In CVPR, 2019.*

Qualitative results

- Image-to-image translation and recurrent baseline networks

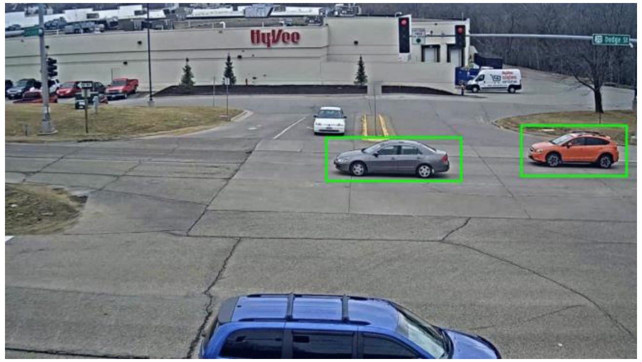


1. *T.-C. Wang et al., High-resolution image synthesis and semantic manipulation with conditional gans. In CVPR, 2018.*
2. *W. Lotter et al., Deep predictive coding networks for video prediction and unsupervised learning. In arXiv preprint arXiv:1605.08104, 2016.*

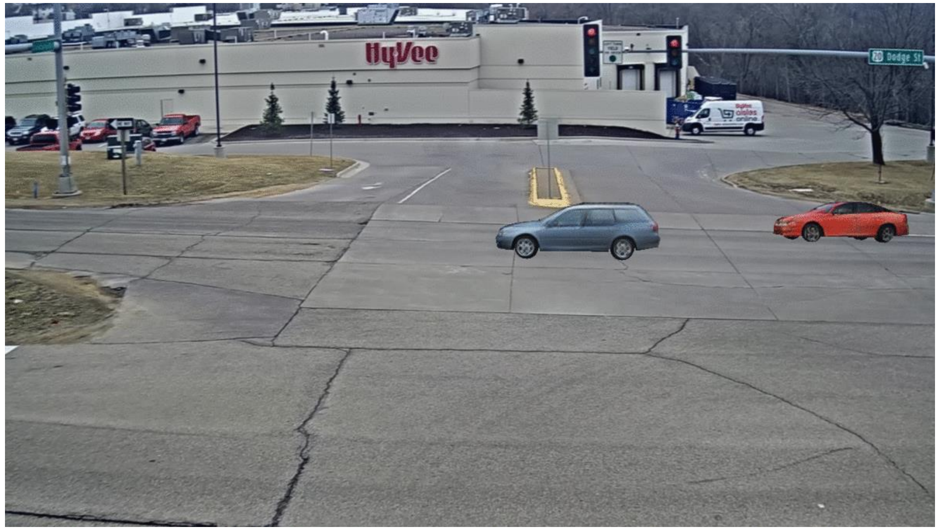
Qualitative results

- Our approach

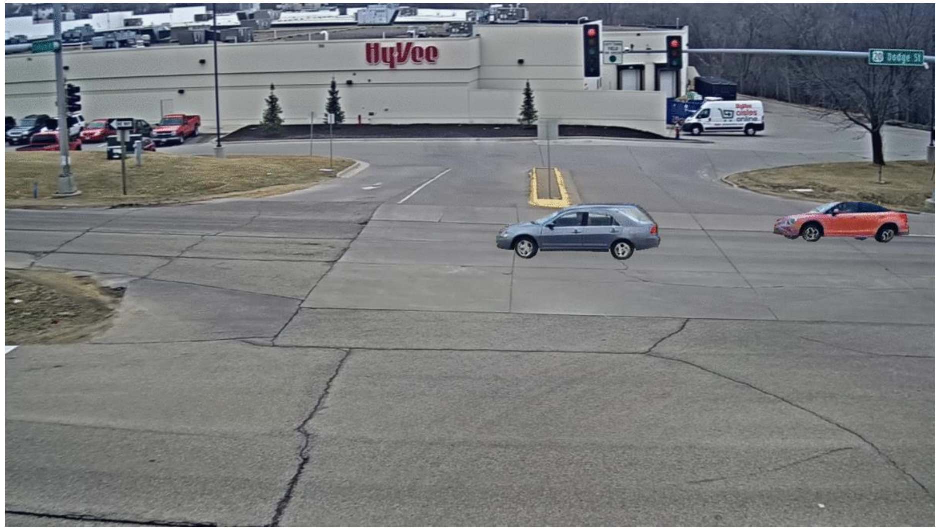
GT scene[1]



VUnet[2] synthetized scene



Warp&Learn[3] synthetized scene



1. Z. Tang et al., *Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In CVPR, 2019.*
2. A. Palazzi et al., *Warp and learn: Novel views generation for vehicles and other objects. In TPAMI, 2020.*
3. P. Esser et al., *A variational u-net for conditional appearance and shape generation. In CVPR, 2018.*

AImage Lab

Quantitative results

- 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)

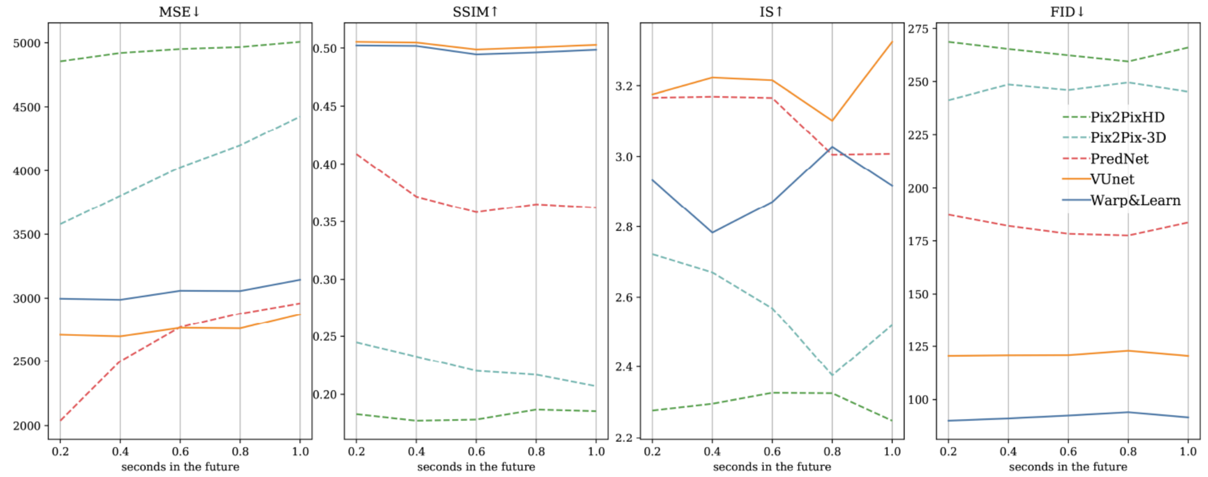


TABLE I

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 [6]	4854	4919	4950	4966	5007
Pix2Pix-3D	3579	3802	4026	4198	4424
PredNet [16]	**2037**	**2499**	2765	2877	2959
Our(VUnet [9])	2705	2692	**2759**	**2755**	**2870**
Our(Warp&Learn [10])	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 [6]	0.18	0.17	0.17	0.18	0.18
Pix2Pix-3D	0.24	0.23	0.22	0.21	0.20
PredNet [16]	0.40	0.37	0.35	0.36	0.36
Our(VUnet [9])	**0.50**	**0.50**	**0.49**	**0.50**	**0.50**
Our(Warp&Learn [10])	**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 [6]	2.27	2.29	2.32	2.32	2.24
Pix2Pix-3D	2.72	2.67	2.56	2.37	2.51
PredNet [16]	3.16	3.16	3.16	3.00	3.00
Our(VUnet [9])	**3.17**	**3.22**	**3.21**	**3.10**	**3.32**
Our(Warp&Learn [10])	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 [6]	274.2	268.6	265.3	262.3	259.4
Pix2Pix-3D	240.6	241.2	248.6	245.9	249.5
PredNet [16]	197.1	197.2	196.4	193.4	196.3
Our(VUnet [9])	192.8	187.3	182.0	178.3	177.49
Our(Warp&Learn [10])	**90.4**	**90.22**	**91.2**	**92.6**	**94.1**

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

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

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

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

Centro Interdipartimentale di Ricerca
Softech: ICT per le Imprese