DeepBEV: A Conditional Adversarial Network for Bird's Eye View Generation

Helmi Fraser, Sen Wang

Perception and Robotics Group, Heriot-Watt University, Edinburgh Centre for Robotics

Abstract

Obtaining a meaningful, interpretable yet compact representation of the immediate surroundings of an autonomous vehicle is paramount for effective operation as well as safety. This work proposes a solution to this by representing semantically important objects from a top-down, ego-centric *bird's eye view* (BEV). The novelty in this work is from formulating this problem as an adversarial learning task, tasking a generator model to produce bird's eye view representations which are plausible enough to be mistaken as a ground truth sample.

This is achieved by using a Wasserstein Generative Adversarial Network based model conditioned on object detections from monocular RGB images and the corresponding bounding boxes. Extensive experiments show our model is more robust to novel data compared to strictly supervised benchmark models, while being a fraction of the size of the next best.

Method

We draw inspiration from recent work in guided image generation, synthesis and super-resolution. Instead of solving the problem in a supervised manner by minimising the difference between the output and ground truth, we instead formulate the problem within the context of adversarial learning.

We seek to leverage the generative capabilities of a Generative Adversarial Network (GAN), specifically a Wasserstein GAN (WGAN) with gradient penalty (WGAN-GP). Our model is composed of two sub-networks: a generator network and a critic network. The generator network is tasked with producing BEV representations from an image, while the critic network is designed to assign a "realness" score to this representation, distinguishing a generated BEV representation from its ground truth counterpart. Therefore, the BEV representation produced by the generator network is gradually trained to be similar to the ground truth.

Model

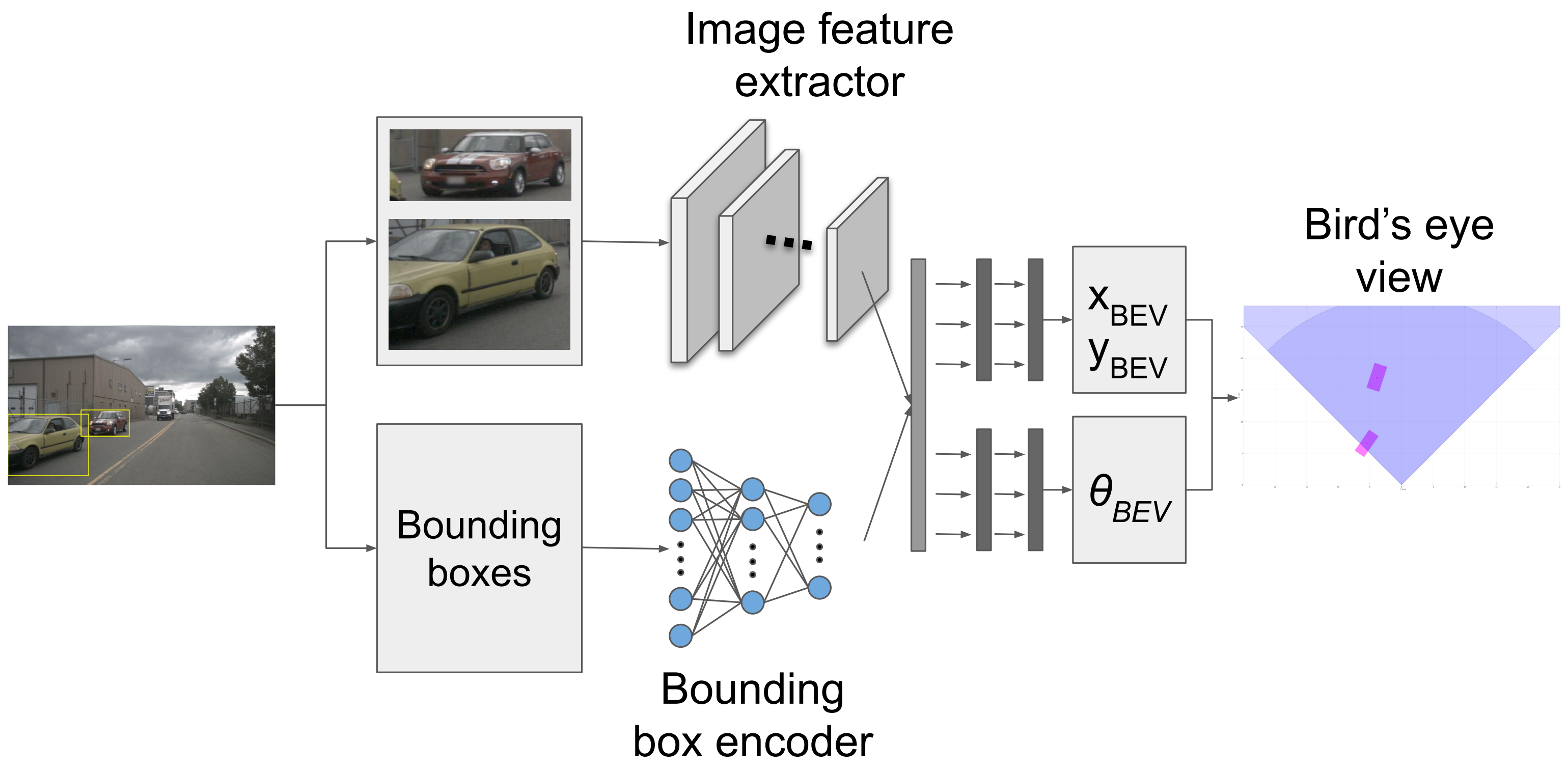

Figure 1 **Generator** system diagram. As input, it accepts object image crops and normalised bounding box co-ordinates and outputs bird's eye view co-ordinates.

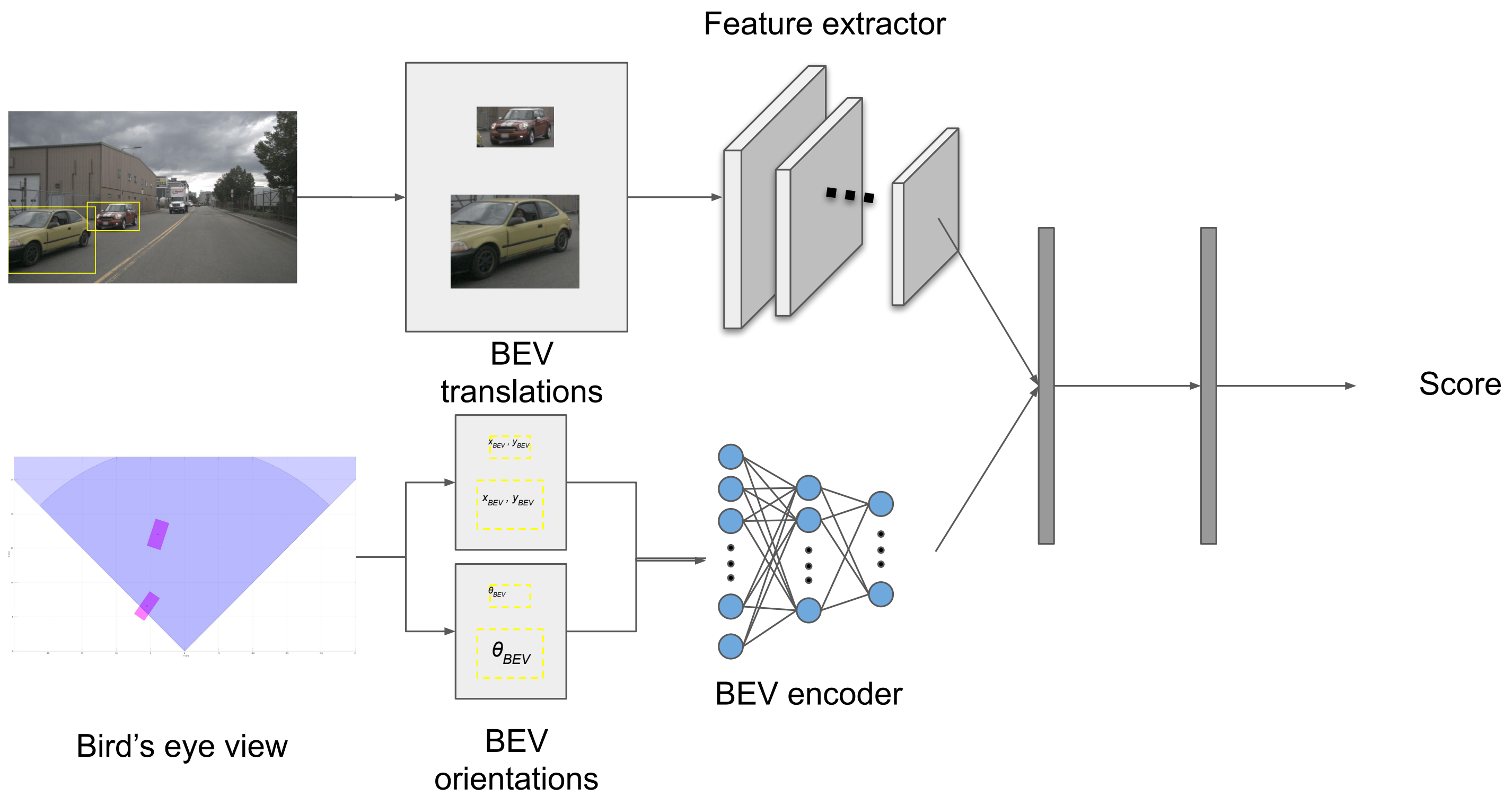

Figure 2 **Critic** system diagram. The critic takes the generated bird's eye view co-ordinates and the image used to generate it, and outputs a coresponding plausibility score.

Results

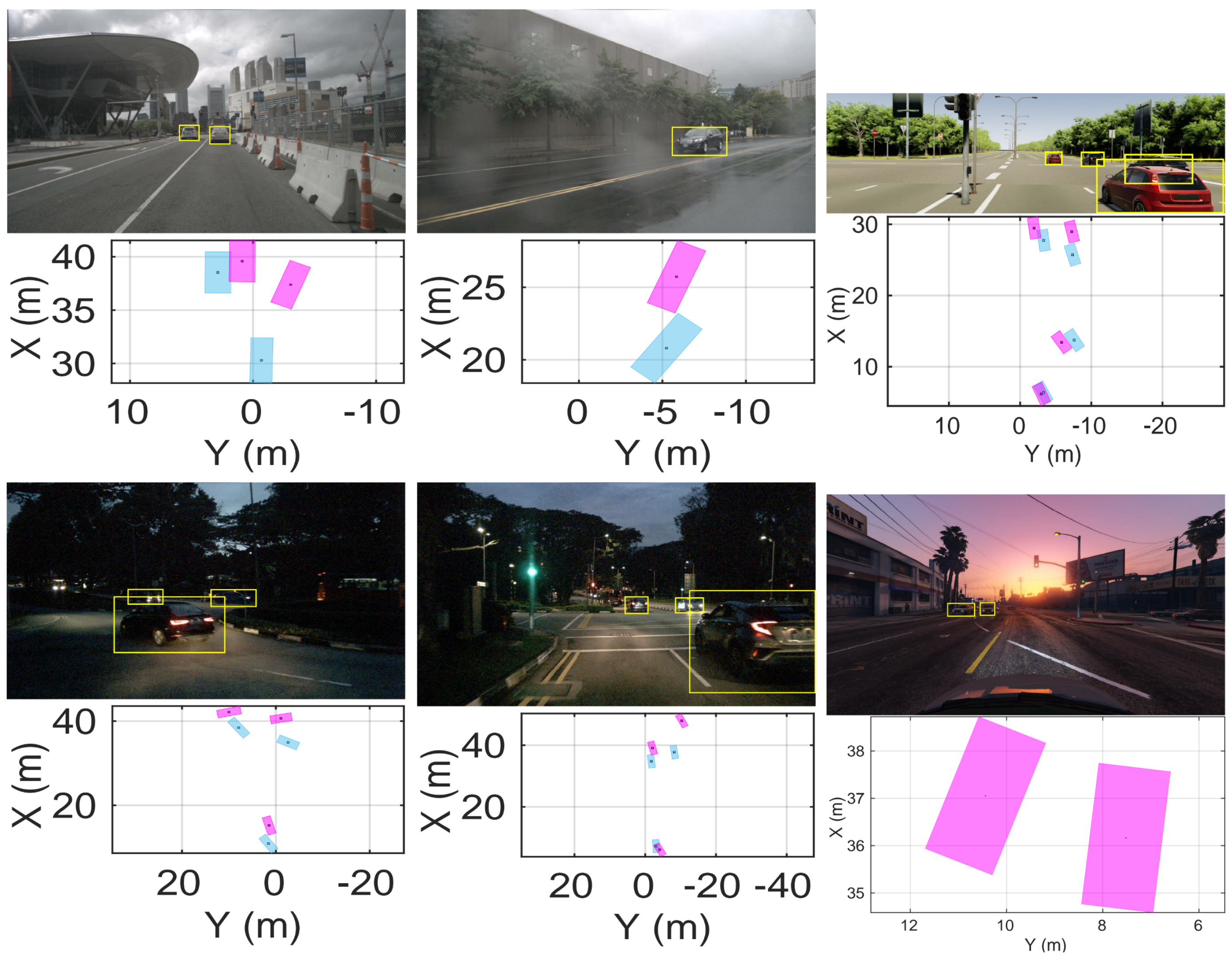

Figure 3 Quantitative and qualitative results. First two collumns are from NuScenes, top right from Virtual KITTI, bottom right from Surround Vehicle Awareness (GTA V). Blue denotes ground truth pose, magenta denotes model prediction.

Results

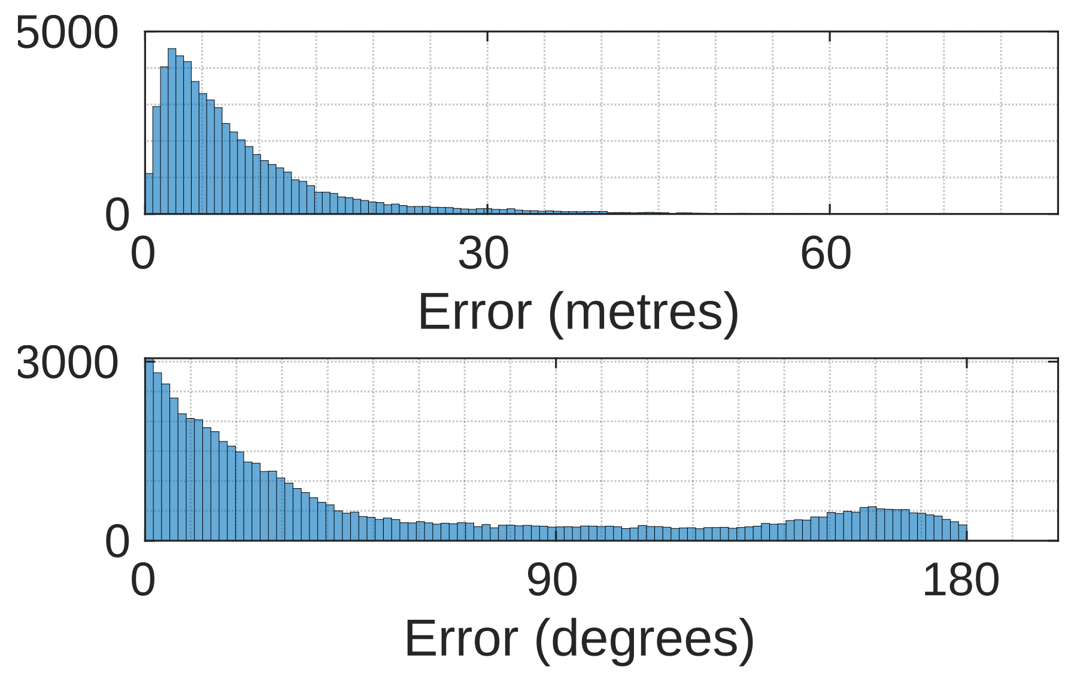

Figure 4 Histogram of DeepBEV translation and rotation errors on NuScenes.

Model	Distance Error (m)		Orientation Error (degrees)	
	Median	SD	Median	SD
DeepBEV	**5.91**	8.22	**28.67**	56.83
ResNet-18	8.62	7.11	30.70	57.40
ResNet-50	8.43	8.44	33.74	57.99
ResNet-101	7.58	9.24	**28.36**	59.29
ResNeXt-50	7.26	8.69	31.86	58.45
Wide ResNet-50	7.97	8.55	33.09	59.08

Table 1 Distance and Orientation Errors on NuScenes.

Acknowledgements

This research is supported by the Engineering and Physical Sciences Research Council (EPSRC), as part of the CDT in Robotics and Autonomous Systems at Heriot-Watt University and The University of Edinburgh.

Contact Details
E-mail: `hmf30@hw.ac.uk`
LinkedIn: helmifraser
Website: https://www.edinburgh-robotics.org/students/helmi-fraser