NIKOLINA CRNOGORAC [1], ALEXANDER SING [1], CSABA BELEZNAI [1], MARKUS VINCZE [2]

[1] AIT Austrian Institute of Technology GmbH, Center for Vision, Automation & Control, Assistive & Autonomous Systems, Giefinggasse 4, 1210 Vienna, Austria. // csaba.beleznai@ait.ac.at //

[2] Vienna University of Technology, Automation and Control Institute, Gußhausstraße 27-29, Vienna, Austria. // vincze@acin.tuwien.ac.at //

MONOCULAR VISION-BASED 3D POSE ESTIMATION FOR ENHANCED CYCLIST SAFETY

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

ABSTRACT

Recent developments in Deep Learning demonstrate that accurately regressing 3D pose parameters from a monocular view is a feasible task. An estimated pose of specific objects with known dimensions from a mobile observer's viewpoint reveals relevant spatial relationship, contributing to an understanding of the surrounding environment. Therefore, monocular 3D pose estimation is an important enabler in safety-related task domains such as perception for autonomous driving and automated traffic monitoring. In our work we present conceptual considerations, a baseline methodology, and results towards monocular vision-based 3D pose estimation involving the safe interaction between cyclists and other vehicles. Furthermore, we propose an enhancement of cyclist detection via learning pose-annotated appearances from our dataset with retro-reflective stripes mounted on the bicycle frame.

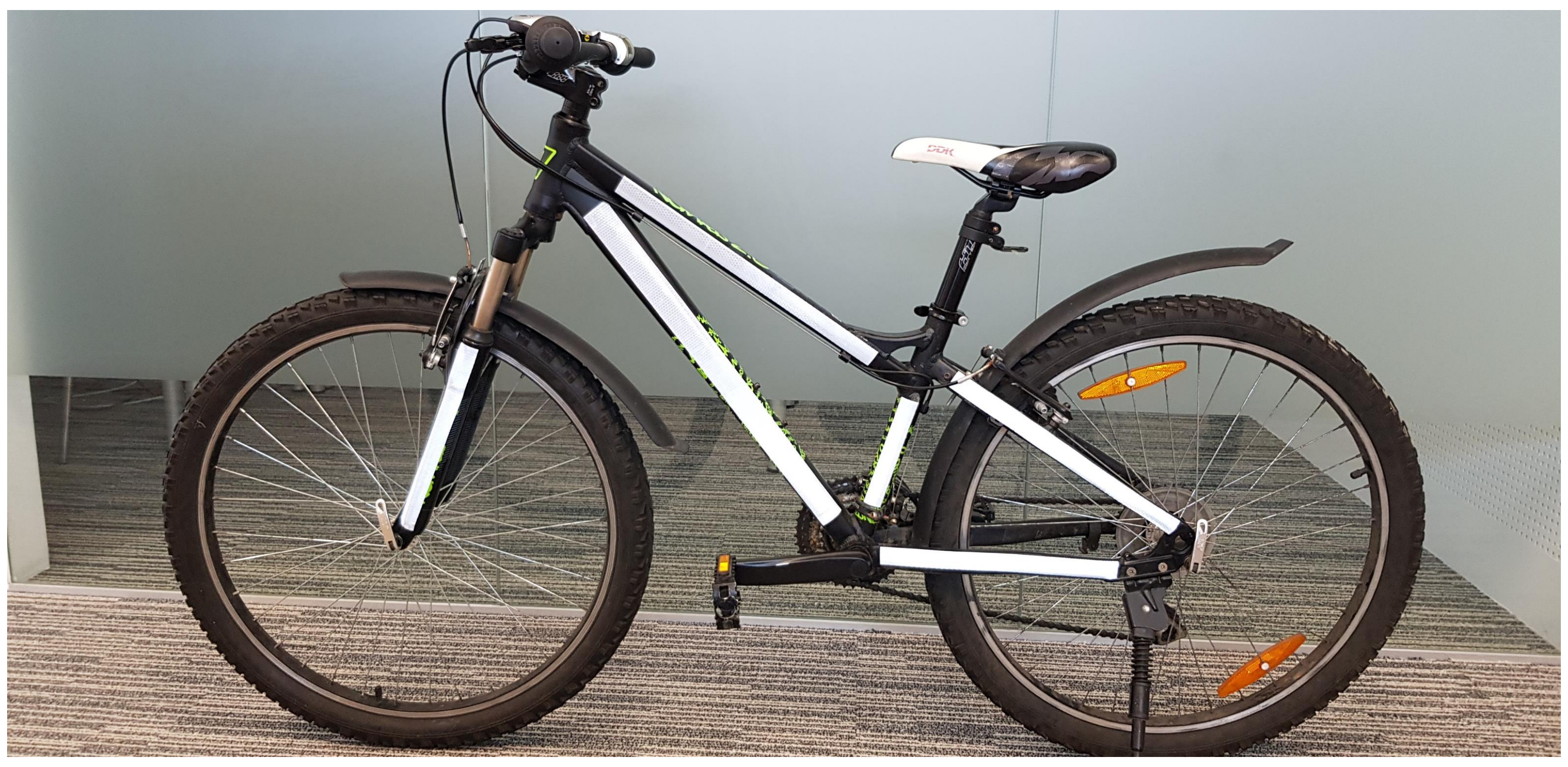

Fig 2. Bicycle equipped with proposed retro-reflective pattern

SOTA

This surge of development of 3D pose estimation in street scenarios partially stems from the emergence of pose-annotated datasets (initiated by the KITTI Vision Benchmark), partially due to the wider use of depth-sensing sensor modalities (stereo vision, LiDAR, Radar) which can derive pose-annotations in an automated manner. Direct learning of the spatial transform from image to BEV space is removing problems associated with scale variation, such as in Pseudo-LIDAR (Wang, 2019) and OF-Transform (Roddick, 2019). In order to to support 3D pose estimation via learning numerous open data-sets like KITTI 3D, Cityscapes 3D, nuScenes etc. have been proposed.

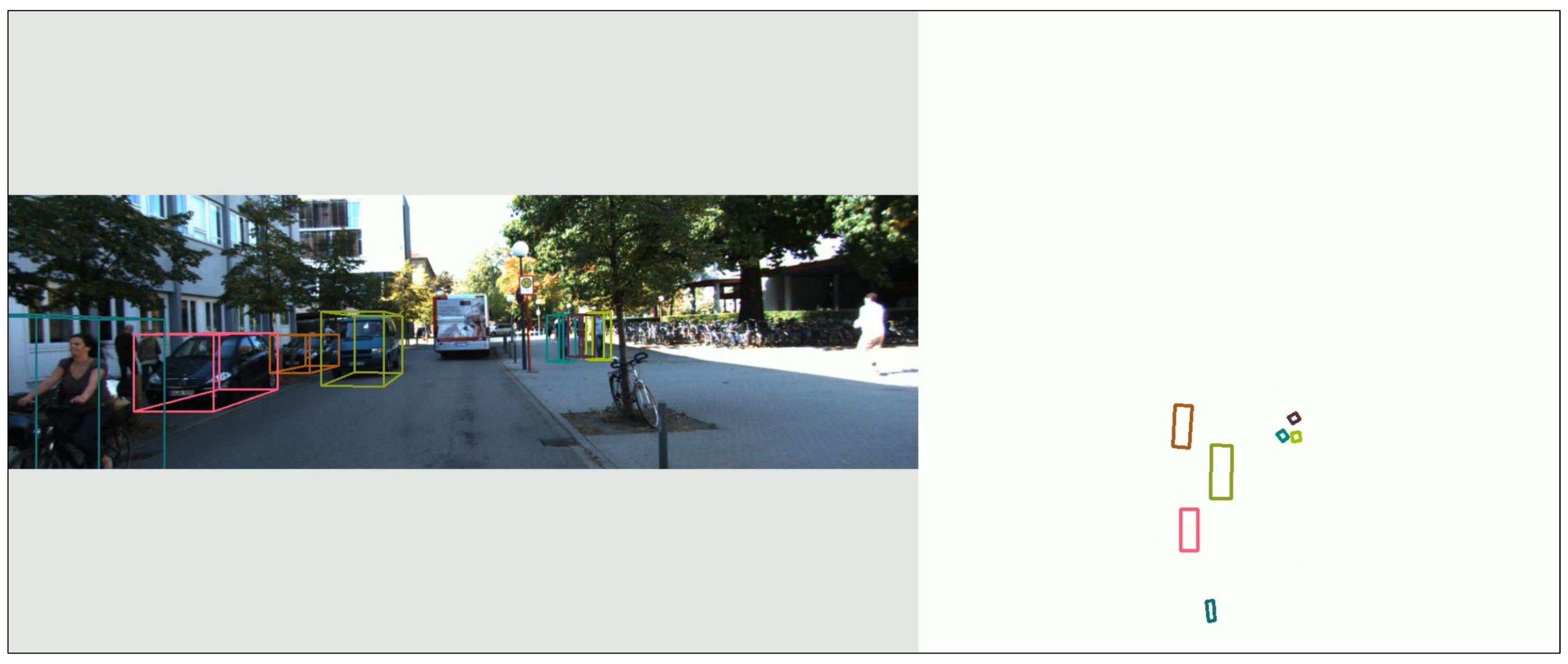

Fig 4. Additional Birds Eye View

RESULTS

Initial results are shown in Fig. 3 and 4 on the KITTI 3D test set. Near-range (<80*m*) detection results are already quite promising, which is essential for cyclist safety. However, detection accuracy, especially orientation estimation, suffers when the distance increases. Therefore, mitigating this issue has been chosen as a research goal, to further enhance the estimates via a large multi-class dataset with diverse spatial scaling.

SUMMARY

Here, we present first results for the pose-aware cyclist and vehicle detection task, with the goal of increasing cyclist safety through automated assessment of traffic.

The next steps will be to elaborate a spatio-temporal attention based multiple object tracking scheme which partitions detection output into consistent trajectories. This shall increase detection accuracy and temporal stability and hence safety.

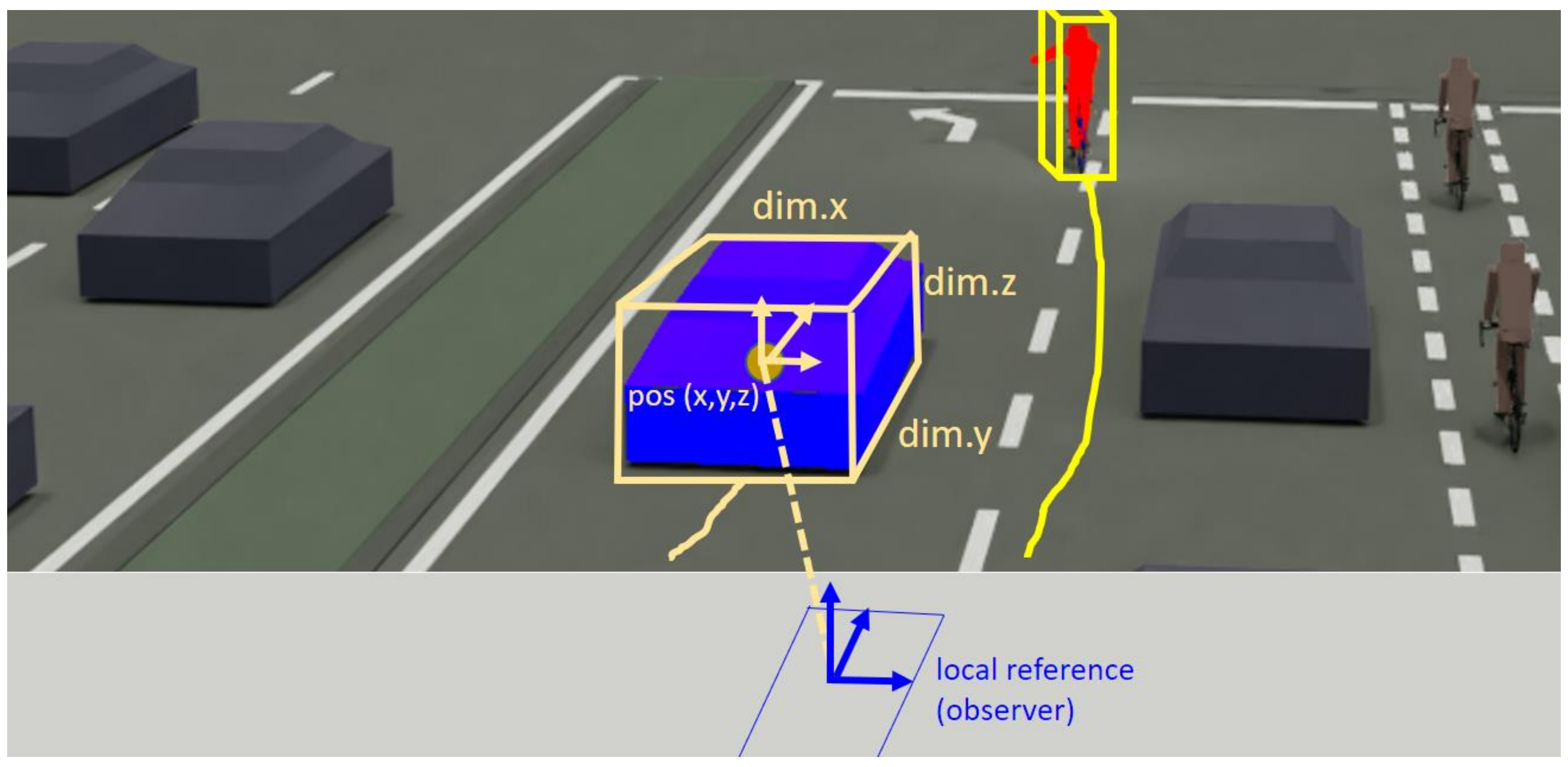

Fig 1. Schematic view of a possible scenario

INTRODUCTION

Our proposed research endeavour relates to automated visual environment perception in complex and dynamic environments.

To enhance the detection and pose estimation accuracy, we propose the generation of a large mixed cyclist dataset, containing pose-annotated real and synthetic images, as well as images captured under adverse photometric conditions, where a specific retro-reflective pattern on the bicycle frame shall generate a visible structure encoding distance and pose, even at poor visibility conditions, such as low-light, fog and heavy rain.

Even with a reduced set of parameters for street-level observations (z = 0, yaw is only relevant object orientation angle), the object distance from the camera is a sensitive parameter to be estimated. This monocular ambiguity can be lowered both by a diverse and accurately 3D-annotated dataset and by algorithmic means, choosing parametric representations which can be more accurately regressed and mapped to a 3D world (birds-eye-view /BEV/) representation frame.

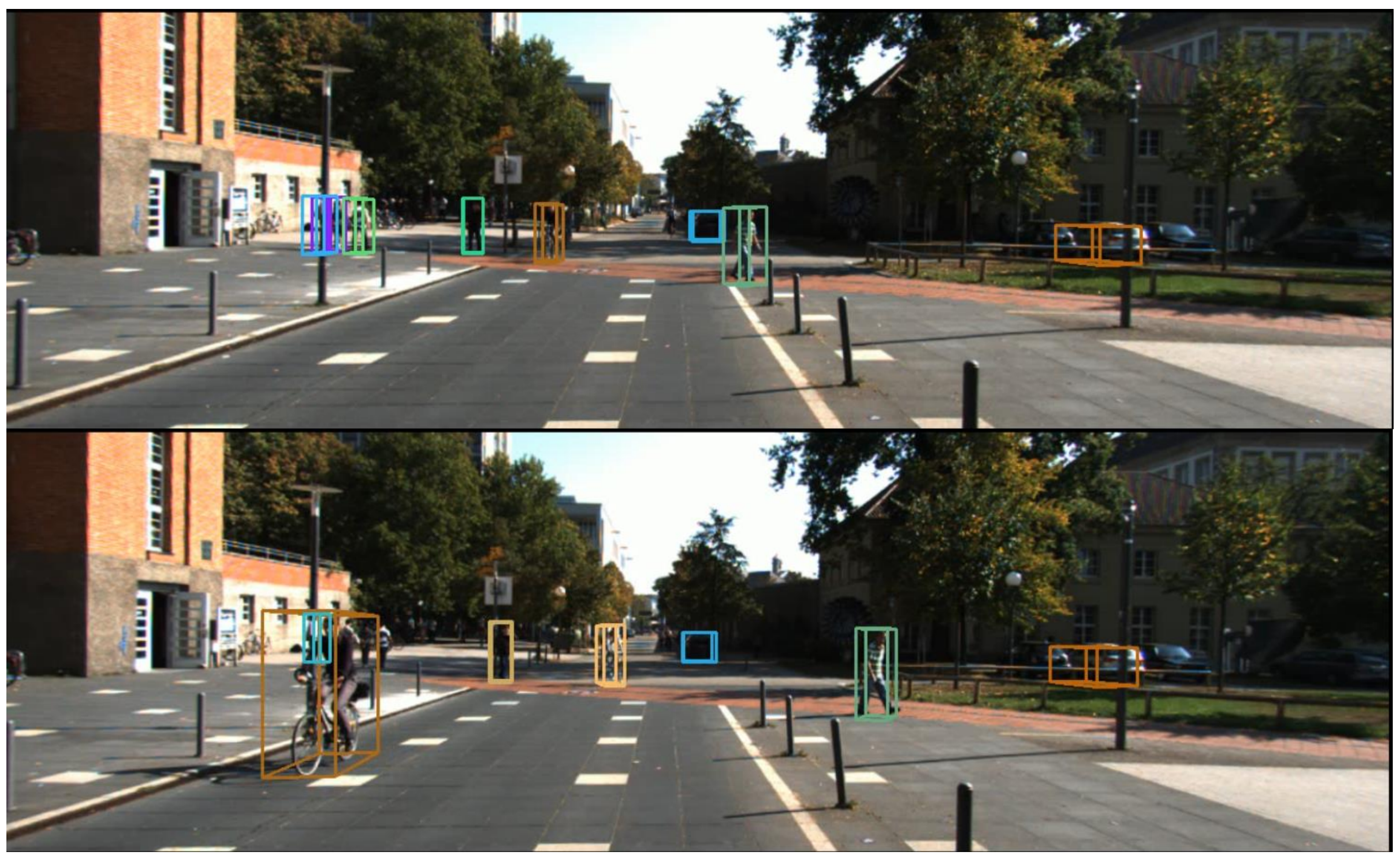

Fig 3. Monocular pose estimation and tracking results using our baseline

METHODOLOGY

Our learning objectives are based on flexible convolutional encoder-decoder-type network (Zhou, 2019). Using the KITTI 3D dataset (Geiger, 2012), we estimate the parameters of object location in the image, metric depth and the yaw angle α of object orientation, encoded as (cosα,sinα).

Inspired by recent advances in target re-identification, we also incorporate a target association scheme exploiting an estimated low-dimensional appearance representation. Additionally, first 2D experiments were conducted using a Transformer-based approach (Zeng et al, 2021).

For the enhancement of cyclist detection under low light conditions we have devised a retro-reflective pattern design, which is low-cost, easy-to-deploy and exhibits a representational compatibility for learning. Enriching the data set with pose-annotated cyclist instances enables us to use the same learning framework with an extended range of illumination conditions. Using a LIDAR sensor, an automated data acquisition procedure is being done.

*This work was carried out within the Bike2CAV project, which is funded by the Austrian Ministry for Climate Action, Environment, Energy, Innovation and Technology (BMK) under the program "Mobility of the Future" and is managed by the Austrian Research Promotion Agency (FFG)