KI DATA TOOLING
The Data Kit for Automotive AI

Augmentation of Pedestrians in Real World Data

Kevin Rösch, Julian Truetsch,
FZI Forschungszentrum Informatik

Why do we use real data for augmentation?
This work tries to avoid the domain gap between generated and real data. It was shown that trained networks can easily spot generated data [1]. This implies, e.g. an object detection network can easily find systematic differences between real and generated pedestrians. If it finds such a shortcut, the object detection performance wouldn't improve for real pedestrians. The usage of real pedestrians for augmentation hinders the network to learn generated features and forces it to focus on the given task.

Pipeline overview
The images are taken from the Cityscapes Dataset [2]. As shown in Fig.1 the augmentation pipeline needs a source image and a target image. One fitting pedestrian from the source image is copied with the information of the semantic label. It is then inserted in the target image at a position that produces a challenging scenario for pedestrian detection. In this example the pedestrian is placed where the traffic sign should occlude him. To achieve the real occlusion and to remove some copy-paste artifacts additional post processing steps are needed.

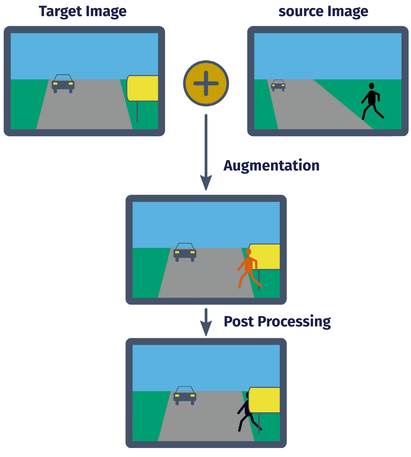

Figure 1: Augmentation Pipeline

Post processing steps
The copy-paste style augmentation pipeline introduces several artifacts that need to be removed to generate a realistic pedestrian in the target image. One of the more advanced problems is to get the occlusion right. Fig. 2 shows the steps needed get the occlusion of the inserted pedestrian correctly. Image a) shows the pedestrian who clearly should be occluded by the car. To get the needed depth for this task, the network MiDaS [3] is used which can predict the depth based on a camera image shown in image b). Fused with the semantic image, that is either given with the labels or predicted with a semantic segmentation network [4] given in image c), we can assign the correct depth value to the inserted pedestrian. It is assumed that the ground is flat. With that assumption the depth value is given by the median depth of the pixels visualized red in image d). As shown in image e) the depth and the semantic masks can be used to cut the pedestrian which results in a realistic occlusion as seen in image f).

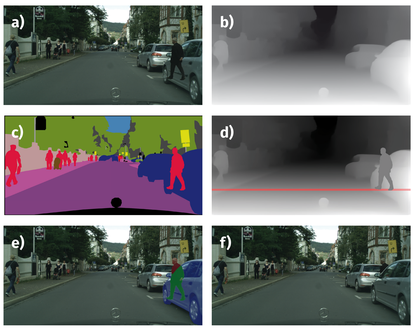

Figure 2: Process to solve the occlusion problem

References:
[1] Wang, Sheng-Yu, et al. „CNN-generated images are surprisingly easy to spot... for now." Proc. Of CVPR. 2020.
[2] Cordts, Marius, et al. „The cityscapes dataset for semantic urban scene understanding." Proc. Of CVPR. 2016.
[3] Ranftl, René, et al. „Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer." arXiv preprint arXiv:1907.01341 (2019).
[4] Romera, Eduardo, et al. „Erfnet: Efficient residual factorized convnet for real-time semantic segmentation." IEEE Transactions on Intelligent Transportation Systems 19.1 (2017):

FZI

For more information contact:
Kevin.roesch@fzi.de

KI Data Tooling is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Climate Action.

KI FAMILIE

Supported by:
Federal Ministry for Economic Affairs and Climate Action
on the basis of a decision by the German Bundestag