Fraunhofer IOSB

FRAUNHOFER INSTITUTE FOR OPTRONICS, SYSTEM TECHNOLOGIES AND IMAGE EXPLOITATION IOSB

IMAGE DOMAIN ADAPTION OF SIMULATED DATA FOR HUMAN POSE ESTIMATION

Thomas Golda, Andreas Blattmann, Jürgen Metzler, Jürgen Beyerer

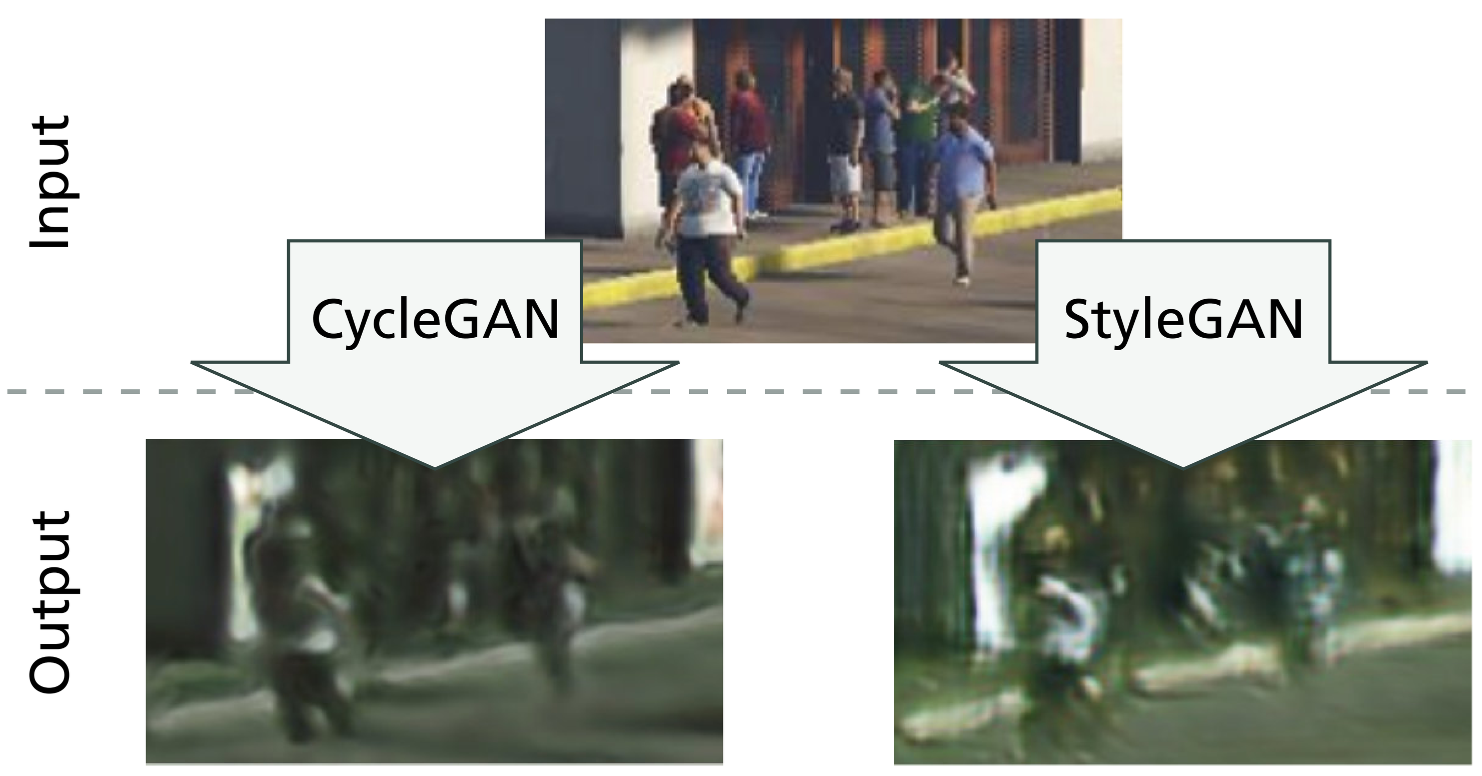
MOTIVATION

- Skeletons are low-dimensional representations of human beings, hence suitable for real-time analysis in 2D computer vision
- Labeled data from COCO and MPII are not suitable since domain gap is quite big
- Synthetic data can be designed in such way to fullfill requirements for application oriented training data
- Pure synthetically generated data comes also with a domain gap that has to be reduced

Person Pose Estimation for Crowded Scenarios

- Training datasets for human pose estimation are not designed for tackling crowded situations
- Earlier attemps like [2] come with a different understanding of „crowdedness“
- Our approach is to generate training data by simulating it using a video game engine as proposed in [1] and adapt this data to our target domain

FIGURE 1



Domain adaption using two different architectures. CycleGAN proves itself with all introduced extensions to provide best results in image detail conservation.

Contribution

- Detailed examination of two different GAN architectures: Cycle-GAN and Style-GAN
- Experiments with various expansions of baseline architecture for detail preservation
- Exhaustive experiments using different target domains plus detailed comparison and evaluation

Experiments

- We evaluated three target domains for the domain adaption task, defined by three datasets: WorldExpo 10 [4], Cityscapes [5] and some internal data
- For evaluation the chosen architecture [3] was trained on the generated data (and on CrowdPose [2])
- Evaluation results are reported on data recorded in a surveillance-like scenario (many people at small scale in urban environment)

RESULTS

Training Dataset	mAP	mAP_{Easy}	AP_{Med}
SyMPose	16.4 ± 0.2	45.1 ± 2.9	16.1 ± 0.3
CrowdPose [2]	23.2 ± 0.4	77.3 ±1.6	22.4 ± 0.3
SyMPose2CS	8.9 ± 0.8	45.9 ± 2.5	8.6 ± 0.9
SyMPose2W10	16.8 ± 0.4	53.2 ± 2.2	16.7 ± 0.5
SyMPose2IOSB	17.4 ± 0.3	49.0 ± 4.3	17.1 ± 0.2
(SyMPose + SyMPose2IOSB)	18.4 ± 0.3	53.0 ± 1.1	17.9 ± 0.3

Conclusion

- Human pose estimation in surveillance applications is very challenging and missing appropriate data
- Adapting synthetically generated data to overcome this lack of data is a way to alleviate the problem
- Finding a suited target domain for the adaption task is crucial for increasing performance

References

[1] M. Fabbri, F. Lanzi, S. Calderara, A. Palazzi, R. Vezzani, and R. Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. European Conference on Computer Vision (ECCV), 2018.

[2] J. Li, C. Wang, H. Zhu, Y. Mao, H. S. Fang, and C. Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[3] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. European Conference on Computer Vision (ECCV), 2018.

[4] C. Zhang, H. Li, X. Wang, and X. Yang. Cross-scene crowd counting via deep convolutional neural networks. Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

Thomas Golda, M. Sc.
Video Exploitation Systems (VID)
thomas.golda@iosb.fraunhofer.de
Phone +49 721 6091-631

KIT
Karlsruhe Institute of Technology