Playing for Data: Ground Truth from Computer Games

Stephan R. Richter\textsuperscript{1,2}, Vibhav Vineet\textsuperscript{1,2}, Stefan Roth\textsuperscript{1}, Vladlen Koltun\textsuperscript{2}
\textsuperscript{1}TU Darmstadt \hspace{1cm} \textsuperscript{2}Intel Labs
\textsuperscript{*Joint first authors}

The Curse of Dataset Annotation
- Recent progress in computer vision is driven by high-capacity models trained on large datasets.
- Current semantic segmentation models appear limited by data rather than capacity.
- Creating large datasets with pixel-level labels requires large amount of human effort.

Idea
- We explore the use of commercial video games for creating large-scale pixel-accurate ground truth data for training semantic segmentation systems.
- Modern open-world games such as Call of Duty, Watch Dogs, and Hitman feature extensive and highly realistic worlds.
- High realism in layout of objects and environments, textures, motion of vehicles and autonomous characters, the presence of small objects that add detail, and the interaction between player and environment.

Extraction
- Challenge: Internal operation and content of off-the-shelf games are largely inaccessible.
- Our solution: We intercept the communication of a game with the underlying graphics library.

Annotation
- Group pixels that share a common combination of mesh, texture, and shader (MTS).
- Annotating MTSs with their semantic class is much faster than outlining objects.
- By recognizing MTSs across frames and game sessions, annotations can be propagated to all other frames of the dataset.

Datset Analysis
- We extracted 24,966 frames with resolution of 1944x1062 pixels from GTAs.
- Labeling 98.9% of their pixels with corresponding semantic classes took just 49 hours.

Evaluation
- Real images
- Synthetic images (all)
- Synthetic images (half)
- Controlled experiments on the CamVid dataset. Training with the full CamVid training set augmented by the synthetic images increases the mean IoU by 5.9 percentage points. Synthetic images also allow reducing the amount of labeled real-world training data by a factor of 3.

References

Acknowledgements
SR1 was supported in part by the German research foundation DFG within the SFB 678. Additional SR1 and TVE were supported in part by the European Commission within the (FP7-2007-2013) ERC Grant Agreement No. 307049. Work on this project was conducted jointly within SR1 and SR2.