Overview

We extend the popular task of multi-object tracking to multi-object tracking and segmentation (MOTS). We provide pixel-level annotations for KITTI and MOTChallenge containing 65,213 pixel masks for 107 objects (cars and pedestrians) in 10,870 video frames. We propose a new baseline method to address detection, tracking, and segmentation with a single convolutional network. We make our annotations, code, and models available at https://www.vision.rwth-aachen.de/page/mots.

Motivation

• Multi-object tracking has been considered mostly on bounding box level
• Bounding box information is often too coarse
• In order to move to the pixel level, we need new datasets and methods

Annotiations

• Semi-automatic annotation procedure applied to KITTI [1] and MOTChallenge [7]

Evaluation Measures

• ρ: (long) sequence from hypothesis to ground truth
• TP: true positives, TN: true negatives
• FP: false positives, FN: false negatives
• M: set of ground truth segmentation masks
• MOTSA: Multi-Object Tracking and Segmentation Accuracy
• MOTSP: Multi-Object Tracking and Segmentation Precision
• sMOTSA: Soft-Multi-Object Tracking and Segmentation Accuracy

Qualitative Results

• Results on KITTI MOTS
• Results on MOTChallenge

Quantitative Results

• Results on KITTI MOTS
• Results on MOTChallenge

Ablations

• Temporal component of TrackR-CNN

More Qualitative Results

• More Qualitative Results

Conclusion

• We’ve introduced MOTS: new task, new annotations, metrics, and baseline
• Training benefits from time-consistent instance segmentations compared to – Single image instance segmentations
• Box-based tracking data
• Be the first to beat our baseline!
• Get the new annotations now at https://www.vision.rwth-aachen.de/page/mots
• KITTI MOTS test-set evaluation server coming soon!

References