MOTS: Multi-Object Tracking and Segmentation

Paul Voigtlaender[1] Michael Krause[1] Aljoša Ošep[1] Jonathon Luiten[1]
Berin Balachandar Gnana Sekar[1] Andreas Geiger[2] Bastian Leibe[1]
[1]RWTH Aachen University [2]MPI for Intelligent Systems and University of Tübingen
{voigtlaender,osep,luiten,leibe}@vision.rwth-aachen.de {michael.krause,berin.gnana}@rwth-aachen.de andreas.geiger@tue.mpg.de

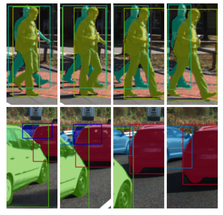
CVPR LONG BEACH CALIFORNIA June 16-20, 2019

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 comprising 65,213 pixel masks for 977 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



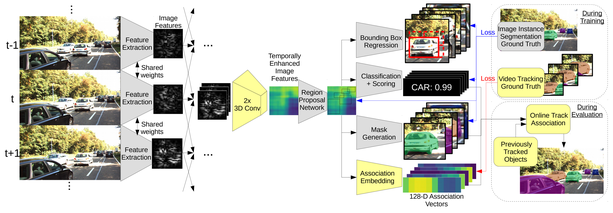
Annotations

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

	KITTI MOTS train	KITTI MOTS val	MOTSChallenge
# Sequences	12	9	4
# Frames	5,027	2,981	2,862
# Tracks Pedestrian	99	68	228
# Masks Pedestrian			
Total	8,073	3,347	26,894
Manually annotated	1,312	647	3,930
# Tracks Car	431	151	-
# Masks Car			
Total	18,831	8,068	-
Manually annotated	1,509	593	-

TrackR-CNN

- Detection, segmentation, and data association in a single convolutional network
- Extends Mask R-CNN by 3D convolutions and association head



- Batch-hard triplet loss [3] for association head

$$\frac{1}{|D|}\sum_{d\in D}\max\Big(\max_{\substack{e\in\mathcal{D}:\\ id_e=id_d}}\|a_e-a_d\|-\min_{\substack{e\in\mathcal{D}:\\ id_e\neq id_d}}\|a_e-a_d\|+\alpha,0\Big).$$

- Learned embeddings for associating detections over time using Euclidean distance

- Associate detections using Hungarian matching (very simple)

Evaluation Measures

- c: (unique) mapping from hypotheses to ground truth
- TP: true positives, $\widetilde{\text{TP}}$: soft number of true positives
- FN: false negatives, FP: false positives, IDS: ID switches
- 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

$$\widetilde{\text{TP}}=\sum_{h\in TP}\text{IoU}(h,c(h))$$

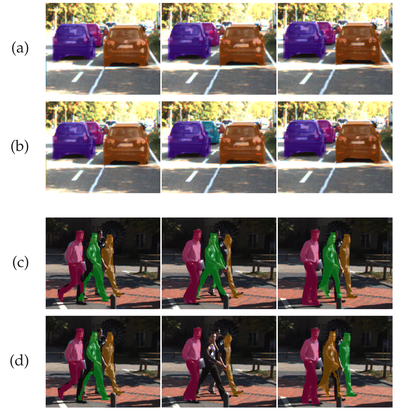
$$\text{MOTSA}=1-\frac{|FN|+|FP|+|IDS|}{|M|}=\frac{|TP|-|FP|-|IDS|}{|M|}$$

$$\text{MOTSP}=\frac{\widetilde{TP}}{|TP|}$$

$$\text{sMOTSA}=\frac{\widetilde{TP}-|FP|-|IDS|}{|M|}$$

Qualitative Results

- (a), (c): TrackR-CNN trained with segmentation masks on KITTI MOTS
- (b), (d): TrackR-CNN trained with boxes only + mask generation by Mask R-CNN



Quantitative Results

- Results on KITTI MOTS

	sMOTSA Car	sMOTSA Ped	MOTSA Car	MOTSA Ped	MOTSP Car	MOTSP Ped
TrackR-CNN (ours)	76.2	46.8	87.8	65.1	**87.2**	**75.7**
Mask R-CNN + maskprop	75.1	45.0	86.6	63.5	87.1	75.6
TrackR-CNN (box orig) + MG	75.0	41.2	87.0	57.9	86.8	76.3
TrackR-CNN (ours) + MG	76.2	**47.1**	87.8	**65.5**	**87.2**	**75.7**
CAMOT [9] (our det)	67.4	39.5	78.6	57.6	86.5	73.1
CIWT [8] (our det) + MG	68.1	42.9	79.4	61.0	86.7	75.7
BeyondPixels [10] + MG	**76.9**	-	**89.7**	-	86.5	-
GT Boxes (orig) + MG	77.3	36.5	90.4	55.7	86.3	75.3
GT Boxes (tight) + MG	**82.5**	**50.0**	**95.3**	**71.1**	**86.9**	**75.4**

  - Line 2: Training using instance segmentation data only
  - Line 3: Training using box based tracking data with post-hoc mask generation
  - +MG: mask generation from bounding boxes using Mask R-CNN
- Results on MOTSChallenge

	sMOTSA	MOTSA	MOTSP
TrackR-CNN (ours)	**52.7**	**66.9**	**80.2**
MHT-DAM [5] + MG	48.0	62.7	79.8
FWT [2] + MG	49.3	64.0	79.7
MOTDT [6] + MG	47.8	61.1	80.0
jCC [4] + MG	48.3	63.0	79.9
GT Boxes (tight) + MG	**55.8**	**74.5**	**78.6**

- MOTS is hard, even when given perfect ground truth bounding boxes!

Ablations

- Temporal component of TrackR-CNN

Temporal component	sMOTSA Car	sMOTSA Ped	MOTSA Car	MOTSA Ped	MOTSP Car	MOTSP Ped
1xConv3D	76.1	46.3	87.8	64.5	87.1	**75.7**
2xConv3D	76.2	**46.8**	87.8	**65.1**	87.2	**75.7**
1xConvLSTM	75.7	45.0	87.3	63.4	87.2	75.6
2xConvLSTM	76.1	44.8	**87.9**	63.3	87.0	75.2
None	**76.4**	44.8	**87.9**	63.2	**87.3**	75.5

- Association mechanism of TrackR-CNN

Association Mechanism	sMOTSA Car	sMOTSA Ped	MOTSA Car	MOTSA Ped	MOTSP Car	MOTSP Ped
Association head	**76.2**	**46.8**	**87.8**	**65.1**	**87.2**	**75.7**
Mask IoU	75.5	46.1	87.1	64.4	**87.2**	**75.7**
Mask IoU (train w/o assoc.)	74.9	44.9	86.5	63.3	87.1	75.6
Bbox IoU	75.4	45.9	87.0	64.3	**87.2**	**75.7**
Bbox Center	74.3	43.3	86.0	61.7	**87.2**	**75.7**

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

[1] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012.
[2] Roberto Henschel, Laura Leal-TaixÃ, Daniel Cremers, and Bodo Rosenhahn. Fusion of head and full-body detectors for multi-object tracking. In *CVPRW*, 2018.
[3] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.
[4] Margret Keuper, Siyu Tang, Bjoern Andres, Thomas Brox, and Bernt Schiele. Motion segmentation & multiple object tracking by correlation co-clustering. *PAMI*, 2018.
[5] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M. Rehg. Multiple hypothesis tracking revisited. In *ICCV*, 2015.
[6] Chen Long, Ai Haizhou, Zhuang Zijie, and Shang Chong. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *ICME*, 2018.
[7] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.
[8] Aljoša Ošep, Wolfgang Mehner, Markus Mathias, and Bastian Leibe. Combined image- and world-space tracking in traffic scenes. In *ICRA*, 2017.
[9] Aljoša Ošep, Wolfgang Mehner, Paul Voigtlaender, and Bastian Leibe. Track, then decide: Category-agnostic vision-based multi-object tracking. *ICRA*, 2018.
[10] S. Sharma, J. A. Ansari, J. K. Murthy, and K. M. Krishna. Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking. In *ICRA*, 2018.