# Beyond Detection: Towards Multi-Object Tracking and Segmentation


Andreas Geiger

Autonomous Vision Group
University of Tübingen / MPI for Intelligent Systems


June 17, 2018


University of Tübingen
MPI for Intelligent Systems
**Autonomous Vision Group**

# MOTS: Multi-Object Tracking and Segmentation

[Voigtlaender, Krause, Osep, Luiten, Sekar, Geiger & Leibe, CVPR 2019]





# Motivation

- Datasets for **multi-object tracking**
  - MOTChallenges
    - MOT15 [Leal-Taixe et al., 2015]
    - MOT16, MOT17 [Milan et al., 2016]
    - CVPR19 [Dendorfer et al., 2019]
  - KITTI Tracking [Geiger et al., 2012]
  - VisDrone2018 [Zhu et al., 2018]
  - DukeMTMC [Ristani et al., 2016]
  - UA-DETRAC [Wen et al., 2015]
  - ...

# Motivation

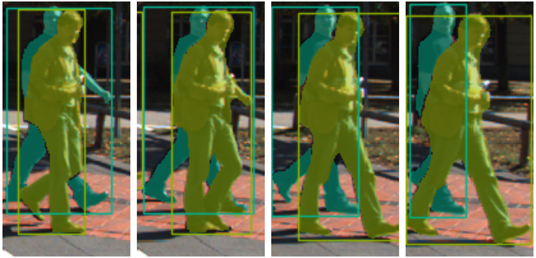
- Datasets for **multi-object tracking**
  - MOTChallenges
    - MOT15 [Leal-Taixe et al., 2015]
    - MOT16, MOT17 [Milan et al., 2016]
    - CVPR19 [Dendorfer et al., 2019]
  - KITTI Tracking [Geiger et al., 2012]
  - VisDrone2018 [Zhu et al., 2018]
  - DukeMTMC [Ristani et al., 2016]
  - UA-DETRAC [Wen et al., 2015]
  - ...
- Led to **great progress** in the community

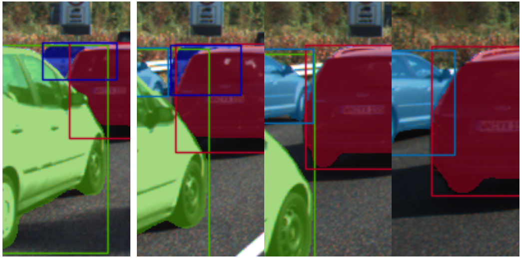
# Motivation

- Datasets for **multi-object tracking**
  - MOTChallenges
    - MOT15 [Leal-Taixe et al., 2015]
    - MOT16, MOT17 [Milan et al., 2016]
    - CVPR19 [Dendorfer et al., 2019]
  - KITTI Tracking [Geiger et al., 2012]
  - VisDrone2018 [Zhu et al., 2018]
  - DukeMTMC [Ristani et al., 2016]
  - UA-DETRAC [Wen et al., 2015]
  - ...
- Led to **great progress** in the community
- But annotations are only on the **bounding box** level

Are bounding boxes enough?

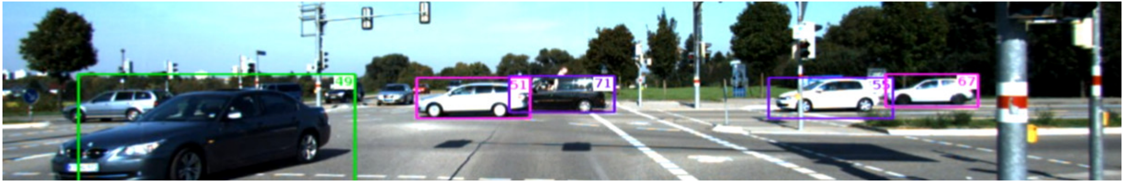
# Object Tracking vs. Segmentation





- In difficult cases, bounding boxes are a very **coarse approximation**
- **Most pixels** of the bounding box **belong to other objects**

# Two Communities




Object Tracking

Semantic Segmentation / Instance Segmentation

Can we unite the two?

# MOTS: Multi-Object Tracking and Segmentation

- Dense pixel-wise annotations are tedious, hard work .. but **we did it!**


KITTI MOTS

# MOTS: Multi-Object Tracking and Segmentation

- Dense pixel-wise annotations are tedious, hard work .. but **we did it!**


MOTSChallenge

# MOTS: Multi-Object Tracking and Segmentation

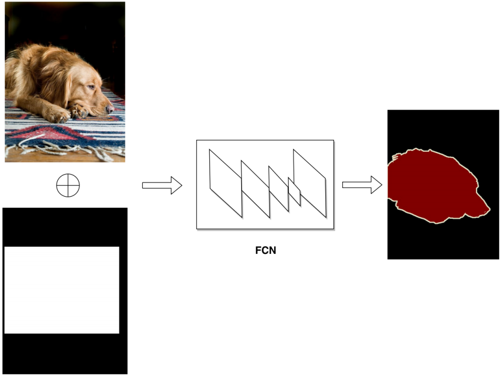
- How? **4 student** assistants & **semi-automatic annotation** procedure

| | KITTI MOTS | | MOTSChallenge |
|---|---|---|---|
| | train | val | train |
| # 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 |
| # Masks Pedestrian (annot.) | 1,312 | 647 | 3,930 |
| # Tracks Car | 431 | 151 | - |
| # Masks Car (total) | 18,831 | 8,068 | - |
| # Masks Car (annot.) | 1,509 | 593 | - |

# Data Annotation

# Data Annotation

- **Starting point:** existing box level tracking annotations
- Fully convolutional network **converts bounding boxes to segmentation masks**



# Data Annotation

- **Starting point:** existing box level tracking annotations
- Fully convolutional network **converts bounding boxes to segmentation masks**
- First, **2 instances** per track are manually annotated

# Data Annotation

- **Starting point:** existing box level tracking annotations
- Fully convolutional network **converts bounding boxes to segmentation masks**
- First, **2 instances** per track are manually annotated
- However, the trained segmentation model will not be perfect

# Data Annotation

- **Starting point:** existing box level tracking annotations
- Fully convolutional network **converts bounding boxes to segmentation masks**
- First, **2 instances** per track are manually annotated
- However, the trained segmentation model will not be perfect
- Repeat until annotations are good:
    1. Annotators **fix worst errors** with polygon annotations
    2. **Add new annotations** to training set of FCN
    3. **Re-train FCN** (pre-train on all, fine-tune per object)
       $\Rightarrow$ Allows for adaptation to appearance and context of each object
    4. **Re-generate masks** using FCN

# Data Annotation

- Manual corrections ensure **consistency** and **high quality**

# Data Annotation

- Manual corrections ensure **consistency** and **high quality**
- Large **savings in annotation time**
  - KITTI MOTS: only 13% of car boxes / 17% of pedestrian boxes manually annotated
  - MOTSChallenge: 15% of pedestrian boxes manually annotated

# Evaluation Metrics

# Evaluation Metrics

- We consider **mask-based variants** of the **CLEAR MOT** metrics [Bernardin and Stiefelhagen, 2008]

# Evaluation Metrics

- We consider **mask-based variants** of the **CLEAR MOT** metrics [Bernardin and Stiefelhagen, 2008]

- Need to **associate** predictions to ground truth instances

  - **Box-based tracking:** boxes might overlap
  - Requires bi-partite matching

# Evaluation Metrics

- We consider **mask-based variants** of the **CLEAR MOT** metrics [Bernardin and Stiefelhagen, 2008]
- Need to **associate** predictions to ground truth instances
  - **Box-based tracking:** boxes might overlap
  - Requires bi-partite matching

  - **Mask-based tracking:** masks are disjoint
  - Establishing correspondences is greatly simplified
  - Hypothesized and ground truth masks are matched iff mask IoU $> 0.5$

# Evaluation Metrics

**(Soft) Multi-Object Tracking and Segmentation Accuracy / Precision:**

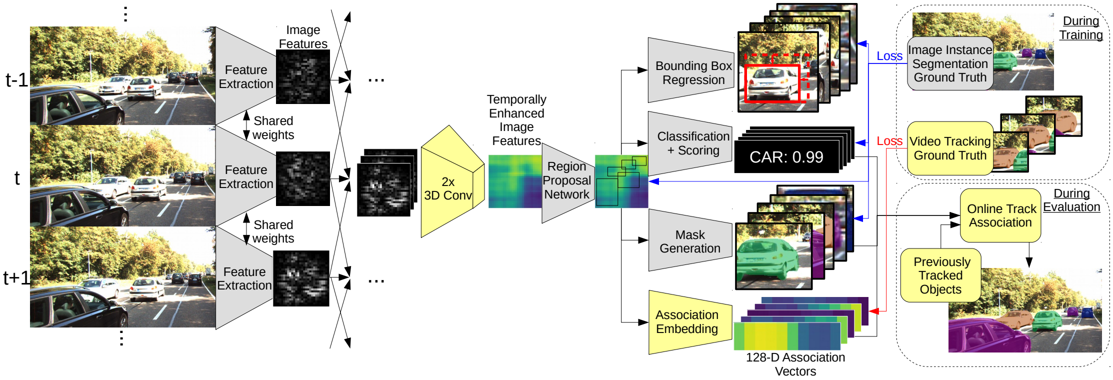
$$\text{MOTSA} = 1 - \frac{|FN| + |FP| + |IDS|}{|M|} = \frac{|TP| - |FP| - |IDS|}{|M|}$$

$$\text{MOTSP} = \frac{\widetilde{TP}}{|TP|} \qquad \text{sMOTSA} = \frac{\widetilde{TP} - |FP| - |IDS|}{|M|} \qquad \widetilde{\text{TP}} = \sum_{h \in TP} \text{IoU}(h, c(h))$$

- $c$: 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

# TrackR-CNN Baseline

# TrackR-CNN




**Key Idea:**

- Detection, segmentation, and data association with a **single ConvNet**
- **Extend Mask R-CNN** by 3D convolutions and association head

# TrackR-CNN

**Association Head:**

- Predict **association vector** for each detection

# TrackR-CNN

**Association Head:**

- Predict **association vector** for each detection
- Detections of same instance should be **close in embedding space**

# TrackR-CNN

**Association Head:**

- Predict **association vector** for each detection
- Detections of same instance should be **close in embedding space**
- Detections of distinct instances should be distant from each other

# TrackR-CNN

**Training:**

- Learned using **batch-hard triplet loss** [Hermans et al., 2017]:

$$\frac{1}{|D|}\sum_{d\in\mathcal{D}}\max\Big(\max_{\substack{e\in\mathcal{D}:\\ id_e=id_d}}\|a_e-a_d\|_2-\min_{\substack{e\in\mathcal{D}:\\ id_e\neq id_d}}\|a_e-a_d\|_2+\alpha,0\Big)$$

- **Mini-batch:** 8 consecutive frames
- **Mine** furthest detection of same instance and closest detection of other instance
- Require separation by not more than **margin** $\alpha$

# TrackR-CNN

**Training:**

- Learned using **batch-hard triplet loss** [Hermans et al., 2017]:

$$\frac{1}{|D|}\sum_{d\in\mathcal{D}}\max\Big(\max_{\substack{e\in\mathcal{D}:\\ id_e=id_d}}\|a_e-a_d\|_2-\min_{\substack{e\in\mathcal{D}:\\ id_e\neq id_d}}\|a_e-a_d\|_2+\alpha,0\Big)$$

- **Mini-batch:** 8 consecutive frames
- **Mine** furthest detection of same instance and closest detection of other instance
- Require separation by not more than **margin** $\alpha$

**Inference:**

- Associate detections over time based on **Euclidean distance** in embedding space and **bi-partite graph matching**

# Experimental Evaluation

# Results of TrackR-CNN on MOTSChallenge


- **Crowded scenes** can lead to **missing detections** and **id switches**

# Results of TrackR-CNN on MOTSChallenge

- **Crowded scenes** can lead to **missing detections** and **id switches**

# Results of TrackR-CNN on MOTSChallenge


- **Crowded scenes** can lead to **missing detections** and **id switches**

# Results of TrackR-CNN on MOTSChallenge


- **Crowded scenes** can lead to **missing detections** and **id switches**

# Results of TrackR-CNN on MOTSChallenge


- **Crowded scenes** can lead to **missing detections** and **id switches**

# Results of TrackR-CNN on MOTSChallenge


- **Crowded scenes** can lead to **missing detections** and **id switches**

# Results of TrackR-CNN on MOTSChallenge

- **Crowded scenes** can lead to **missing detections** and **id switches**

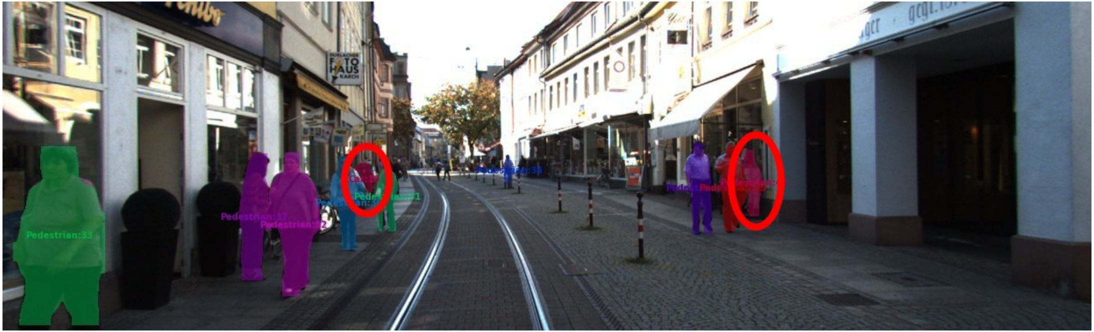
# Results of TrackR-CNN on MOTSChallenge


- **Crowded scenes** can lead to **missing detections** and **id switches**

# Results of TrackR-CNN on KITTI MOTS


- Most objects distinguished well but some **erroneous detections** remain (red)

# Results of TrackR-CNN on KITTI MOTS




- Most objects distinguished well but some **erroneous detections** remain (red)

# Results of TrackR-CNN on KITTI MOTS


- Most objects distinguished well but some **erroneous detections** remain (red)

# Results of TrackR-CNN on KITTI MOTS


- Most objects distinguished well but some **erroneous detections** remain (red)

# Results of TrackR-CNN on KITTI MOTS

- **Continuation of track** with same ID after missing detection (red)

# Results of TrackR-CNN on KITTI MOTS

- **Continuation of track** with same ID after missing detection (red)

# Results of TrackR-CNN on KITTI MOTS

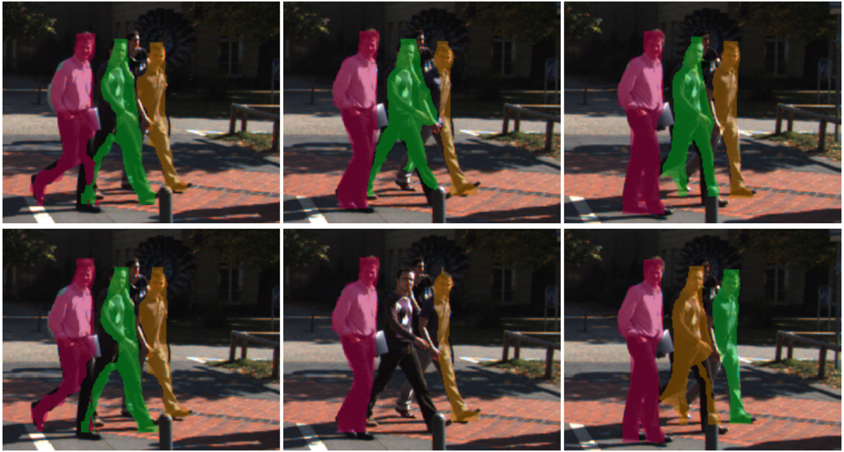
- **Continuation of track** with same ID after missing detection (red)

# Comparison to Box Detection + Mask Prediction

Top: TrackR-CNN    Bottom: TrackR-CNN (box) + Mask R-CNN

- Training with masks **avoids confusion** between similar nearby objects

# Comparison to Box Detection + Mask Prediction



Top: TrackR-CNN    Bottom: TrackR-CNN (box) + Mask R-CNN

- Training with masks **avoids confusion** between similar nearby objects

# Quantitative Results on KITTI MOTS

| | sMOTSA | | MOTSA | | MOTSP | |
|---|---|---|---|---|---|---|
| | Car | Ped | Car | Ped | Car | Ped |
| TrackR-CNN (mask) | **76.2** | **46.8** | **87.8** | **65.1** | **87.2** | **75.7** |
| Mask R-CNN + Optic Flow Propagation | 75.1 | 45.0 | 86.6 | 63.5 | 87.1 | 75.6 |
| TrackR-CNN (box) + Mask R-CNN | 75.0 | 41.2 | 87.0 | 57.9 | 86.8 | 76.3 |
| GT Boxes (orig) + Mask R-CNN | 77.3 | 36.5 | 90.4 | 55.7 | 86.3 | 75.3 |
| GT Boxes (tight) + Mask R-CNN | 82.5 | 50.0 | 95.3 | 71.1 | 86.9 | 75.4 |

- TrackR-CNN **improves over** training on **single instances and box tracks**
- Compared to the flow propagation baseline, our method runs in **real-time**

# Quantitative Results on MOTSChallenge

| | sMOTSA | MOTSA | MOTSP |
|---|---|---|---|
| TrackR-CNN (mask) | **52.7** | **66.9** | **80.2** |
| MHT-DAM [Kim et al., 2015] + Mask R-CNN | 48.0 | 62.7 | 79.8 |
| FWT [Henschel et al., 2018] + Mask R-CNN | 49.3 | 64.0 | 79.7 |
| MOTDT [Long et al., 2018] + Mask R-CNN | 47.8 | 61.1 | 80.0 |
| jCC [Keuper et al., 2018] + Mask R-CNN | 48.3 | 63.0 | 79.9 |
| GT Boxes (tight) + Mask R-CNN | 55.8 | 74.5 | 78.6 |

- **MOTS is challenging** – even with perfect ground truth bounding boxes
- Segmenting pedestrians in **crowded scenes** is difficult

# Ablation Study: Temporal Model on KITTI MOTS

| Temporal component | sMOTSA | | MOTSA | | MOTSP | |
|---|---|---|---|---|---|---|
| | Car | Ped | Car | Ped | Car | 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 |

- **Conv3D improves** for pedestrians, but **ConvLSTM does not**
- But overall **effect is limited** $\rightarrow$ Better ways to incorporate temporal context?

# Ablation Study: Association Mechanism on KITTI MOTS

| Association Mechanism | sMOTSA | | MOTSA | | MOTSP | |
|---|---|---|---|---|---|---|
| | Car | Ped | Car | Ped | Car | 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** |
| 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** |

- Mask IoU: associate based on IoU of mask warped using **optic flow** (PWC-Net)
- Bbox IoU: associate based on bounding box warped using **median optic flow**
- Bbox Center: associate based on **unwarped box center** distance

# More Results

# Summary

- **MOTS:** new **task, annotations, metrics, baselines**

# Summary

- **MOTS:** new **task, annotations, metrics, baselines**
- Training benefits from time-consistent instance segmentations compared to
  - Single image instance segmentations
  - Box-based tracking data

# Summary

- **MOTS:** new **task, annotations, metrics, baselines**
- Training benefits from time-consistent instance segmentations compared to
  - Single image instance segmentations
  - Box-based tracking data
- Be the first to **beat our baseline!**

# Summary

- **MOTS:** new **task, annotations, metrics, baselines**
- Training benefits from time-consistent instance segmentations compared to
  - Single image instance segmentations
  - Box-based tracking data
- Be the first to **beat our baseline!**
- Annotations and code: https://www.vision.rwth-aachen.de/page/mots

# KITTI MOTS Challenge

**The KITTI Vision Benchmark Suite**

A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago

Karlsruhe Institute of Technology

home setup stereo flow sceneflow depth odometry object tracking road semantics raw data submit results

Andreas Geiger (MPI Tübingen) | Philip Lenz (KIT) | Christoph Stiller (KIT) | Raquel Urtasun (University of Toronto)

## Multi-Object Tracking and Segmentation (MOTS) Evaluation

This benchmark is under construction. Currently, you can download the training set of the MOTS benchmark.
The test set and evaluation will be released soon.

- Download training set

**Coming soon:** http://www.cvlibs.net/datasets/kitti/eval_mots.php

Thank you!

http://autonomousvision.github.io

Federal Ministry of Education and Research

intel

Microsoft Research

NVIDIA