STEP: Segmenting and Tracking Every Pixel

Andreas Geiger

Autonomous Vision Group
University of Tübingen and MPI for Intelligent Systems

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

MAX-PLANCK-GESELLSCHAFT

ellis
European Laboratory for Learning and Intelligent Systems

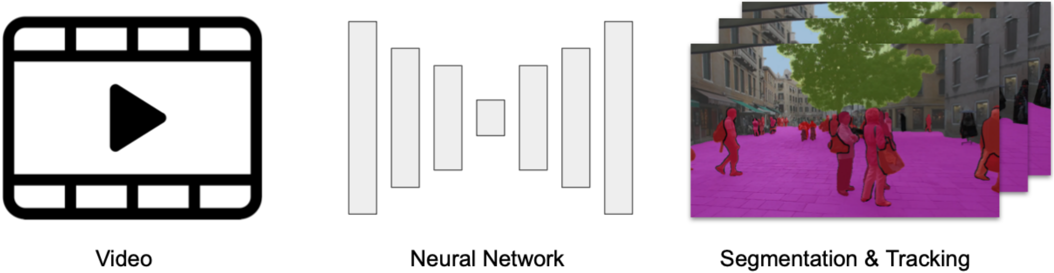
Covered Paper

STEP: Segmenting and Tracking Every Pixel

Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, Aljosa Osep, Laura Leal-Taixe and Liang-Chieh Chen

NeurIPS Track on Datasets and Benchmarks 2021

Overview



- **Task:** Video Panoptic Segmentation
- **Goal:** Assign semantic classes and track identities to all pixels in a video
- **Contribution:** New benchmarks (KITTI-STEP, MOTChallenge-STEP) & new metric

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Why Segmentation matters

- Estimating drivable area
- Semantic understanding of surroundings
- Pixel-precise instance understanding

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Why Tracking matters

- Anticipate the temporal evolution of objects
- Obstacle avoidance
- Motion planning

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Segmenting and Tracking every Pixel

MOTS: No dense segmentation

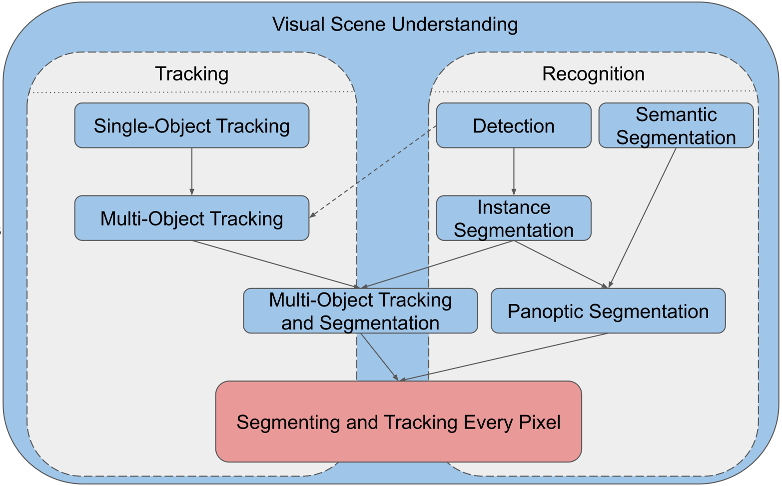
Panoptic Segm.: No tracking

- Existing datasets and benchmarks either lack dense segmentation or tracking
- Our goal: **Segmenting and tracking every pixel (STEP)** for long time periods
- Video Panoptic Segmentation [Kim et al. 2020], but new benchmarks and metrics

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Evolution of Visual Scene Understanding

Are we ready for autonomous driving?



DeepLab

Mask R-CNN

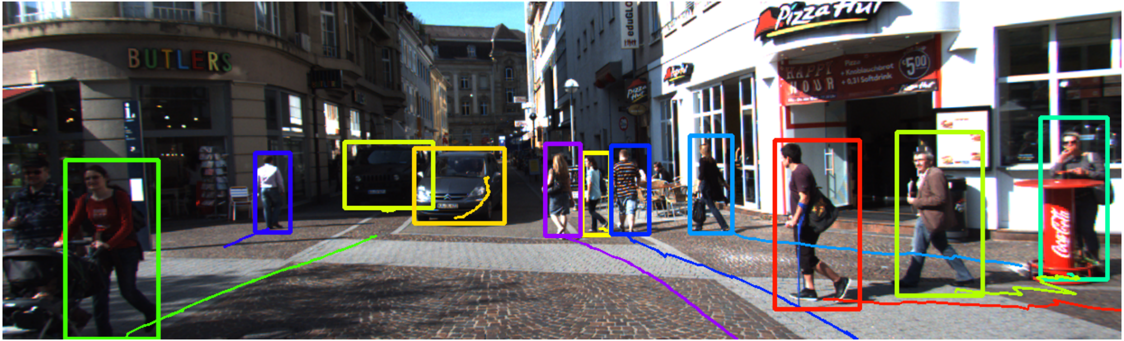
M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

From Tracking to STEP

New: Spatially and temporally dense annotated KITTI-STEP and MOTChallenge-STEP

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

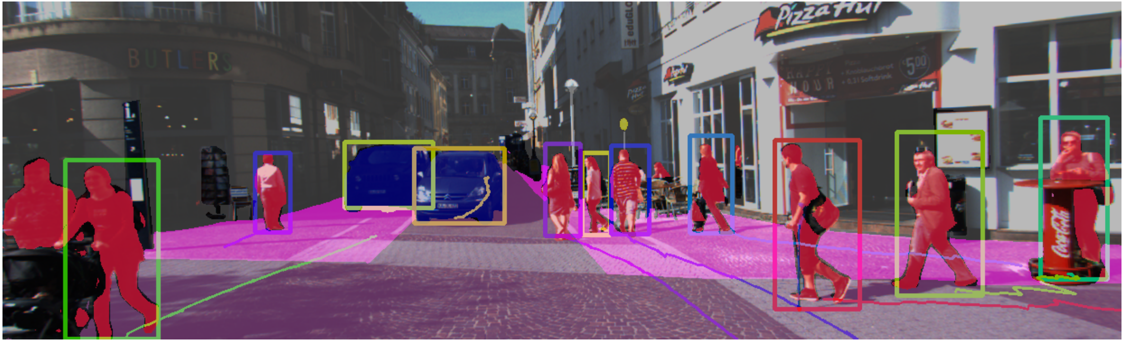
From Tracking to STEP



New: Spatially and temporally dense annotated KITTI-STEP and MOTChallenge-STEP

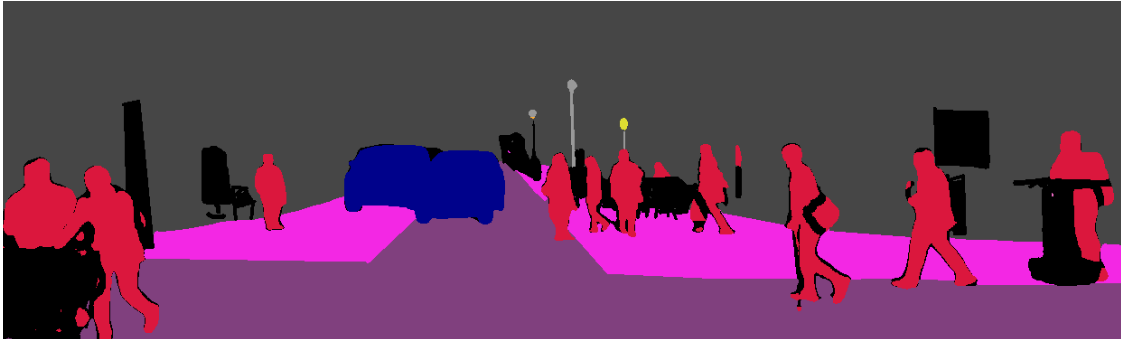
M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

From Tracking to STEP



New: Spatially and temporally dense annotated KITTI-STEP and MOTChallenge-STEP

From Tracking to STEP

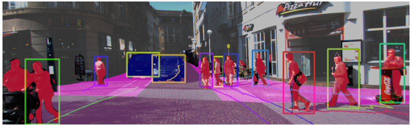


New: Spatially and temporally dense annotated KITTI-STEP and MOTChallenge-STEP

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Contributions

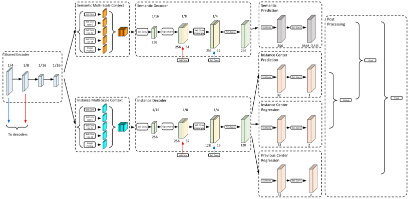
- **KITTI-STEP** and **MOTChallenge-STEP**



- A novel **pixel-centric metric STQ**

$$STQ = (SQ \times AQ)^{\frac{1}{2}}$$

- **Baselines** tackling both segmentation and tracking



KITTI-STEP and MOTChallenge-STEP

Existing Datasets

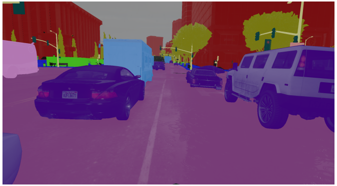
- **KITTI-MOTS** and **MOTSChallenge**

Missing segmentation labels.

- **Cityscapes-VPS**

Every clip has 6 annotated frames (every 5^{th} frame) and spans only 1.8 seconds.

- **Synthetic datasets**

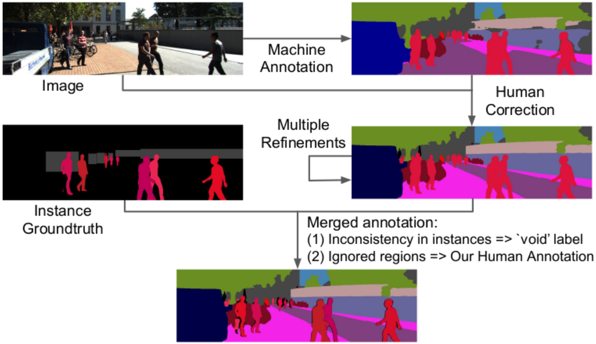


Issues with insufficient photo-realism and thus domain shift.

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Annotation Process

- KITTI-MOTS and MOTSChallenge as basis
- Annotate every frame semi-automatically with semantic segmentation
- Merge tracks and semantic segmentation



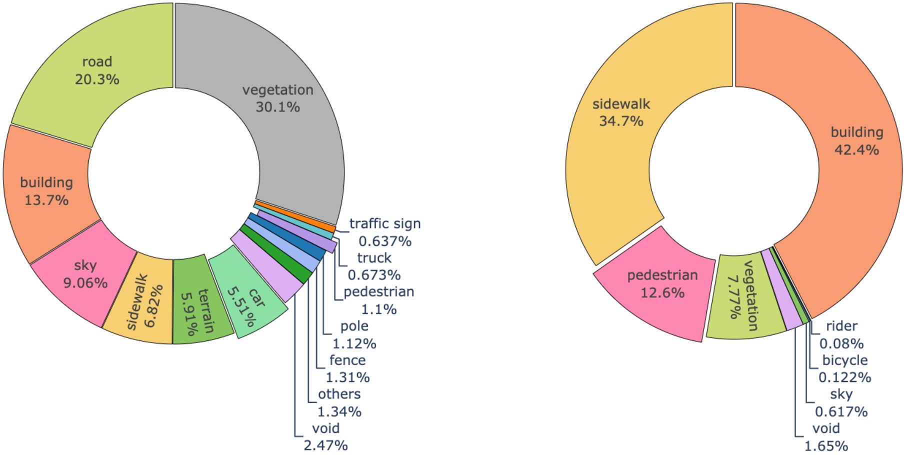
M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Dataset Comparison (Training Set)

Dataset statistics	Cityscapes-VPS	KITTI-STEP	MOTChallenge-STEP
# Sequences (trainval/test)	450 / 50	21 / 29	2 / 2
# Frames (trainval/test)	2,700 / 300	8,008 / 10,173	1,125 / 950
# Tracking classes	8	2	1
# Semantic classes	19	19	7
# Annotated Masks†	72,171	126,529	17,232
Every frame annotated	✗	✓	✓
Annotated frame rate (FPS)	3.4	10	30
	Max/Mean/Min	Max/Mean/Min	Max/Mean/Min
Annotated frames per seq.†	6 / 6 / 6	1,059 / 381 / 78	600 / 562 / 525
Track length (frames)†	**6 / 3 / 1**	**643 / 51 / 1**	**569 / 187 / 1**

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Tracking the most salient classes

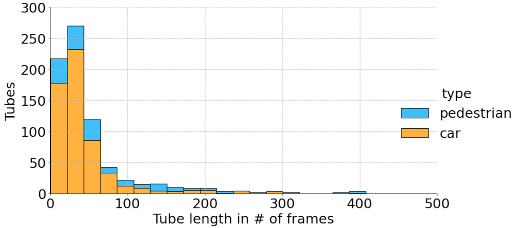


KITTI-STEP

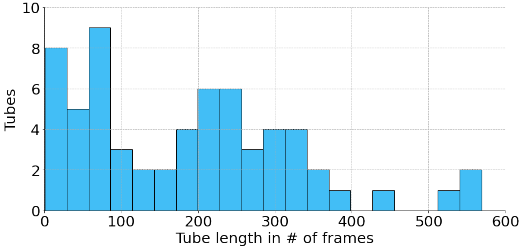
MOTChallenge-STEP

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Track length distribution



KITTI-STEP



MOTChallenge-STEP

- In real-world sequences, tracks last much longer than a few frames
- STEP enables evaluation of **long-term tracking**

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

STQ: Segmentation and Tracking Quality

Why a new metric?

The focus of benchmark papers is usually on the dataset and on baselines.
Few papers do a **thorough analysis** on the metric aspect.

- What is a bad metric?
- What is a good metric?
- What properties do we want?
- Can the metric be tricked?

Existing metrics such as **Video Panoptic Quality** [Kim et al. 2020] and **Panoptic Tracking Quality** [Hurtado et al. 2020] build upon metrics for panoptic segmentation and multi-object tracking, thereby inheriting their drawbacks.

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

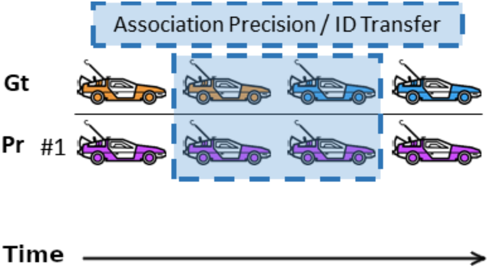
Metric design

Metric Properties	STQ	PTQ	VPQ
P1: Analyze full videos at **pixel level** (not segment level)	✓	✗	(✓)
P2: Avoid **thresholding** (e.g., for TP vs. FP classification)	✓	✗	✗
P3: No penalty for **ID recovery** (correcting mistakes)	✓	✗	✗
P4: Consider **precision** and **recall** for association	✓	✗	(✓)
P5: **Decouple** segmentation and tracking errors	✓	✗	✗

- Panoptic Tracking Quality (PTQ): Penalizes error recovery, negative scores
- Video Panoptic Quality (VPQ): Designed for sparse annotations & short clips

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Metric design



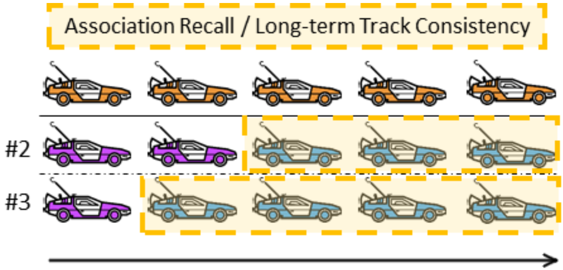
Nr.	**STQ** (↑)	PTQ (↑)	VPQ† (↑)
#1	0.71	1.0	0.0

†VPQ computed on whole sequence.

- STQ is the only metric that properly penalizes ID transfer

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Metric design



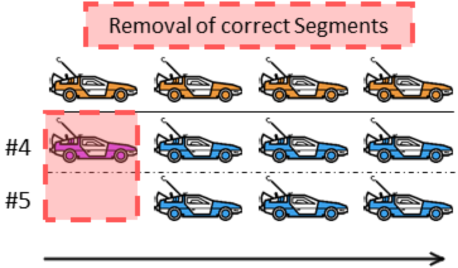
Nr.	**STQ** (↑)	PTQ (↑)	VPQ† (↑)
#2	0.72	0.8	0.4
#3	0.82	0.8	0.53

†VPQ computed on whole sequence.

- STQ and VPQ encourage long-term track consistency

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Metric design



Nr.	**STQ** (↑)	PTQ (↑)	VPQ† (↑)
#4	0.79	0.75	0.5
#5	0.65	0.86	0.75

†VPQ computed on whole sequence.

- Only STQ reduces score when removing semantically correct segments

Formal Definition

The task of **Segmenting and Tracking Every Pixel (STEP)** requires a function

$$f(x, y, t) \mapsto (c, id)$$

which maps every pixel (x, y, t) to a semantic class c and a track ID id.

- We denote the ground-truth as $gt(x, y, t)$ and the prediction as $pr(x, y, t)$
- **STQ** measures **Association Quality (AQ)** and **Segmentation Quality (SQ)**

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Association Quality (AQ)

We define the **prediction** and **ground-truth** for a particular id as follows:

$$pr_{id}(id) = \{(x, y, t) | pr(x, y, t) = (*, id)\}$$
$$gt_{id}(id) = \{(x, y, t) | gt(x, y, t) = (*, id)\}$$

- The proposed AQ is designed to work at a **pixel-level** of a full video (P1)
- All associations have an influence on the score, **no IoU threshold** (P2)
- Semantic segmentation errors are not penalized in AQ (P5), only in SQ

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Association Quality (AQ)

We define the **true positive associations (TPA)** of a specific ID as follows:

$$TPA(p, g) = |pr_{id}(p) \cap gt_{id}(g)|$$

Similarly, false negative associations (FNA) and false positive associations (FPA) can be defined to compute precision P_{id} and recall R_{id}. To account for the effect of both precision and recall (P4), we define the basic building block IoU_{id} for AQ as follows:

$$IoU_{id}(p, g) = \frac{P_{id}(p, g) \times R_{id}(p, g)}{P_{id}(p, g) + R_{id}(p, g) - P_{id}(p, g) \times R_{id}(p, g)}$$

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Association Quality (AQ)

Following our goal of long-term track consistency, we encourage **ID recovery** (P3) by weighting the score of each predicted tube by its TPA. **Association Quality (AQ):**

$$AQ(g) = \frac{1}{|gt_{id}(g)|} \sum_{p, |p \cap g| \neq \emptyset} TPA(p, g) \times IoU_{id}(p, g),$$

$$AQ = \frac{1}{|\mathrm{gt\_tracks}|} \sum_{g \in \mathrm{gt\_tracks}} AQ(g).$$

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Segmentation Quality (SQ)

We use **Intersection-over-Union (IoU)** to measure segmentation quality.
Formally, given $pr(x, y, t)$, $gt(x, y, t)$ and class c we define:

$$pr_{sem}(c) = \{(x, y, t) | pr(x, y, t) = (c, *)\}$$

$$gt_{sem}(c) = \{(x, y, t) | gt(x, y, t) = (c, *)\}$$

We then define the **Segmentation Quality (SQ)** as the mean IoU score:

$$IoU(c) = \frac{|pr_{sem}(c) \cap gt_{sem}(c)|}{|pr_{sem}(c) \cup gt_{sem}(c)|}$$

$$SQ = mIoU = \frac{1}{|\mathbf{C}|} \sum_{c \in \mathbf{C}} IoU(c)$$

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Segmentation and Tracking Quality (STQ)

We combine both scores into **Segmentation and Tracking Quality (STQ)** via the geometric mean:

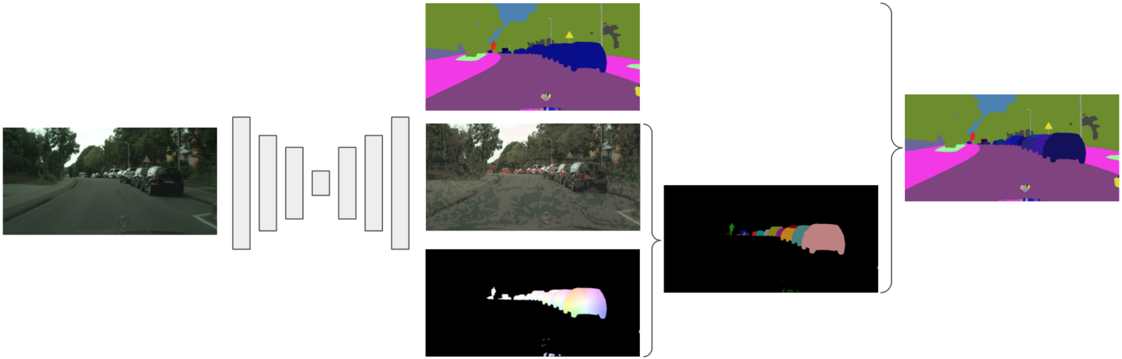
$$STQ = (AQ \times SQ)^{\frac{1}{2}}$$

- STQ hence requires methods to perform well in both segmentation and tracking

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Baselines

Panoptic-DeepLab



Panoptic-DeepLab: [Cheng et al., 2017]

- State-of-the-art per-frame panoptic segmentation network
- 3 branches: semantic segmentation, center heatmap, pixel-to-center offsets.

Extensions to Tracking

Single-frame baselines:

- **B1: IoU Association.** The predicted thing segments of two consecutive frames are matched by Hungarian Matching with a mask IoU threshold $\delta = 0.3$. To account for occluded objects, unmatched predictions are kept for 10 frames.
- **B2: SORT** [Bewley et al., 2016]. Bi-partite matching between sets of Kalman filter track predictions and object detections based on the bounding box overlap.
- **B3: Mask Propagation.** Uses RAFT optical flow [Teed et al., ECCV] to warp each predicted mask at frame $t-1$ into frame t, followed by the IoU matching (B1).

Multi-frame baseline:

- **B4: Center Motion.** Add prediction head to the base network in order to regress every pixel to its instance center in the previous frame. Inspired by CenterTrack.

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Results

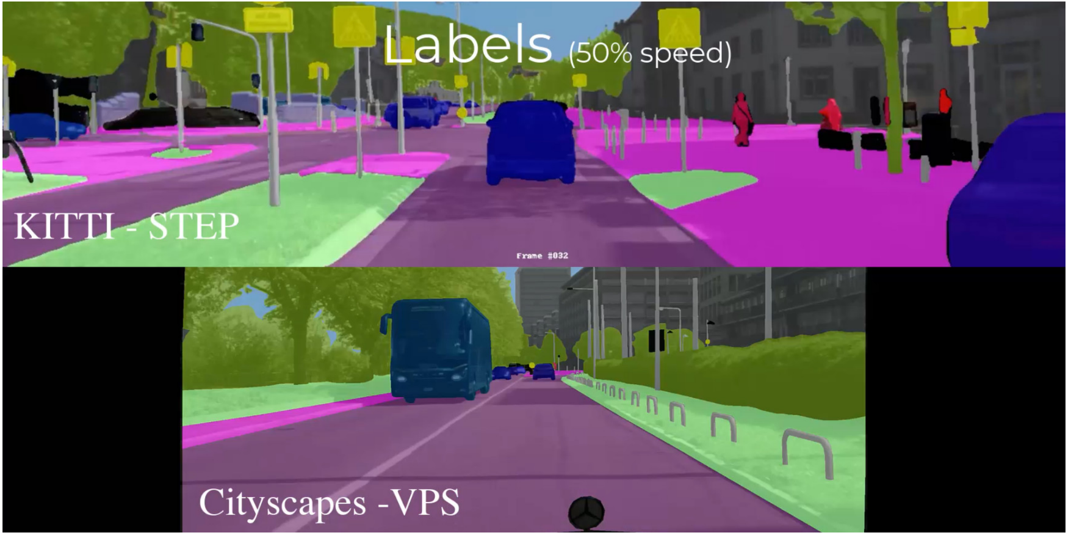
Results on KITTI-STEP

KITTI-STEP	OF	STQ	AQ	SQ	VPQ	PTQ
B1: IoU Association	✗	0.58	0.47	**0.71**	**0.44**	*0.48*
B2: SORT	✗	*0.59*	0.50	**0.71**	0.42	*0.48*
B3: Mask Propagation	✓	**0.67**	**0.63**	**0.71**	**0.44**	**0.49**
B4: Center Motion	✗	0.58	0.51	*0.67*	0.40	0.45
VPSNet (Kim et al.)	✓	0.56	*0.52*	0.61	*0.43*	**0.49**

- Single-frame methods (separate segmentation and tracking) perform best
- Combining SotA segmentation and tracking yields best results (B3)
- More work needed to exploit full potential of multi-frame methods

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Video



M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Resources

- KITTI-STEP: http://www.cvlibs.net/datasets/kitti/eval_step.php
- MOTChallenge-STEP: https://motchallenge.net/data/STEP-ICCV21/
- DeepLab2: https://github.com/google-research/deeplab2
- ICCV 2021 Workshop:

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Summary

- We present a **new perspective** on the task of **video panoptic segmentation**
- We provide a **new benchmark** (STEP) focusing on measuring algorithm performance at the most **detailed** level possible, taking each pixel into account
- Our benchmark and metric are designed for evaluating algorithms in real-world scenarios where understanding **long-term tracking** performance is important
- We believe that this work provides an important STEP towards a **dense, pixel-precise video understanding**

M. Weber et al.: STEP: Segmenting and Tracking Every Pixel. NeurIPS Track on Datasets and Benchmarks, 2021.

Thank you!

http://autonomousvision.github.io

erc
DFG
Federal Ministry of Education and Research
Federal Ministry for Economic Affairs and Energy
intel
Microsoft Research
NVIDIA