Learning Multi-Object Tracking and Segmentation from Automatic Annotations: Supplementary Material

Lorenzo Porzi[†], Markus Hofinger[‡], Idoia Ruiz[*], Joan Serrat[*], Samuel Rota Bulò[†], Peter Kontschieder[†]
Mapillary Research[†], Graz University of Technology[‡], Computer Vision Center, UAB[*]
research@mapillary.com[†], markus.hofinger@icg.tugraz.at[‡], {iruiz,joans}@cvc.uab.es[*]

Abstract

This document contains:

- *A detailed description of the MOT / MOTS metrics*
- *All the hyper-parameters involved in our data generation and network training processes*
- *Experimental results on the MOTSChallenge dataset*
- *More details on signed intersection over union and its use as a cue in tracklet matching*

1. CLEAR MOT and MOTS metrics

The CLEAR MOT metrics (including MOTA and MOTP) are first defined in [1] to evaluate Multi-Object Tracking systems. In order to compute MOTA and MOTP, a matching between the ground truth and predicted tracklets needs to be computed at each frame by solving a linear assignment problem (LAP). We will not repeat the details of this process, instead focusing on the way its outputs are used to compute the metrics. In particular, for each frame t, the matching process gives:

- the number of correctly matched boxes TP_t;
- the number of false positive boxes FP_t, *i.e.* the predicted boxes that are not matched to any ground truth;
- the number of mismatched boxes IDS_t, *i.e.* the boxes belonging to a predicted tracklet that was matched to a different ground truth tracklet in the previous frame;
- the number of ground truth boxes GT_t;
- the intersection over union $\text{IoU}_{t,i}$ between each correctly predicted box and its matching ground truth.

Given these, the metrics are defined as:

$$\text{MOTA} = \frac{\sum_t (\text{TP}_t - \text{FP}_t - \text{IDS}_t)}{\sum_t \text{GT}_t} ,$$

and

$$\text{MOTP} = \frac{\sum_{t,i} \text{IoU}_{t,i}}{\sum_t \text{TP}_t} .$$

The MOTS metrics [7] extend the CLEAR MOT metrics to the segmentation case. Their computation follows the overall procedure described above, with a couple of exceptions. First, box IoU is replaced with mask IoU. Second, the matching process is simplified by defining a ground truth and predicted segment to be matching if and only if their IoU is greater than 0.5. Different from the bounding box case, the segmentation masks are assumed to be non-overlapping, meaning that this criterion results in a unique matching without the need to solve a linear assignment problem (LAP). With these changes, and given the definitions above, the MOTS metrics are:

$$\text{MOTSA} = \frac{\sum_t (\text{TP}_t - \text{FP}_t - \text{IDS}_t)}{\sum_t \text{GT}_t} ,$$

$$\text{sMOTSA} = \frac{\sum_t (\sum_i \text{IoU}_{t,i} - \text{FP}_t - \text{IDS}_t)}{\sum_t \text{GT}_t} ,$$

and

$$\text{MOTSP} = \frac{\sum_{t,i} \text{IoU}_{t,i}}{\sum_t \text{TP}_t} .$$

2. Hyper-parameters

2.1. Data Generation

In the data generation process, when constructing tracklets (see Sec. 3.2 of the main paper) we use the following parameters: $\tau_0 = 10\text{pix}$, $\tau_1 = 10\text{pix}$, $\tau_2 = 2$.

2.2. Network training

As mentioned in Section 5.2 of the main paper, all our trainings follow a linear learning rate schedule:

$$\text{lr}_i = \text{lr}_0 \left(1 - \frac{i}{\#\text{steps}}\right) ,$$

where the initial learning rate lr_0 and total number of steps depend on the dataset and pre-training setting. The actual

Dataset	Pre-training	lr_0	# epochs	N_w
KITTI Synth	I	0.02	20	12
	M	0.01	10	12
KITTI Synth, KITTI MOTS sequences	I	0.02	180	12
KITTI MOTS	I	0.02	180	12
	M, KS	0.01	90	12
BDD100k	all	0.02	100	10
MOTSChallenge	I	0.01	90	10
	C	0.02	180	10

Table 1: MOTSNet training hyperparameters for different datasets and pre-training settings.

Method	Pre-training	sMOTSA	MOTSA	MOTSP
TrackR-CNN [7]	I, C, M	52.7	66.9	80.2
MHT-DAM [4]	I, C, M	48.0	62.7	79.8
FWT [2]	I, C, M	49.3	64.0	79.7
MOTDT [5]	I, C, M	47.8	61.1	80.0
jCC [3]	I, C, M	48.3	63.0	79.9
MOTSNet	I	41.8	55.2	78.4
	I, C	**56.8**	**69.4**	**82.7**

Table 2: Results on the MOTSChallenge dataset. Top section: state-of-the-art results using masks from [7]. Bottom section: our MOTSNet results under different pre-training settings.

values, together with the per-GPU batch sizes N_w are reported in Tab. 1. The loss weight parameter λ in the first equation of Sec. 4.2 of the main paper is fixed to 1 in all experiments, except for the COCO pre-trained experiment on MOTSChallenge, where $\lambda = 0.1$.

3. MOTSNet results on MOTSChallenge

In Tab. 2 we present our results on MOTSChallenge, the second dataset contributed in [7] and again compare against all related works reported therein. This dataset comprises of 4 sequences, a total of 2.862 frames and 228 tracks with roughly 27k pedestrians, and is thus significantly smaller than KITTI MOTS. Due to the smaller size, the evaluation in [7] runs leave-one-out cross validation on a per-sequence basis. We again report numbers for differently pre-trained versions of MOTSNet. The importance of segmentation pre-training on such small datasets is quite evident: while MOTSNet (I) shows the overall worst performance, its COCO pre-trained version significantly improves over all baselines. We conjecture that this is also due to the type of scenes – many sequences are recorded with a static camera and crossing pedestrians are shown in a quite close-up setting (see *e.g*. Fig. 1).

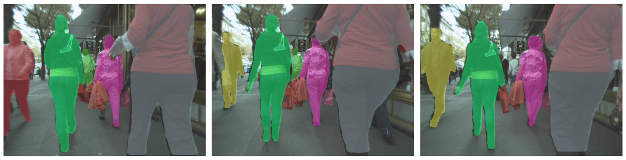

Figure 1: Sample MOTSNet predictions on a sub-sequence from the MOTS Challenge dataset.

4. Signed Intersection over Union

Signed Intersection over Union, as defined in [6], extends standard intersection over union between bounding boxes, by providing meaningful values when the input boxes are not intersecting. Given two bounding boxes $\hat{\boldsymbol{b}} = (\hat{u}_1, \hat{v}_1, \hat{u}_2, \hat{v}_2)$ and $\boldsymbol{b} = (u_1, v_1, u_2, v_2)$, where (u_1, v_1) and (u_2, v_2) are the coordinates of a box's top-left and bottom-right corners, respectively, the signed intersection over union $\text{sIoU}(\hat{\boldsymbol{b}}, \boldsymbol{b})$ is:

- greater than 0 and equal to standard intersection over union when the boxes overlap;
- less than 0 when the boxes don't overlap, and monotonically decreasing as their distance increases.

This is obtained by defining:

$$\text{sIoU}(\hat{\boldsymbol{b}}, \boldsymbol{b}) = \frac{|\hat{\boldsymbol{b}} \sqcap \boldsymbol{b}|_\pm}{|\hat{\boldsymbol{b}}| + |\boldsymbol{b}| - |\hat{\boldsymbol{b}} \sqcap \boldsymbol{b}|_\pm} ,$$

where

$$\hat{\boldsymbol{b}} \sqcap \boldsymbol{b} = \begin{pmatrix} \max(\hat{u}_1, u_1) \\ \max(\hat{v}_1, v_1) \\ \min(\hat{u}_2, u_2) \\ \min(\hat{v}_2, v_2) \end{pmatrix}$$

is an extended intersection operator, $|\boldsymbol{b}|$ denotes the area of $\boldsymbol{b}$, and

$$|\boldsymbol{b}|_\pm = \begin{cases} +|\boldsymbol{b}| & \text{if } u_2 > u_1 \wedge v_2 > v_1, \\ -|\boldsymbol{b}| & \text{otherwise,} \end{cases}$$

is the "signed area" of $\boldsymbol{b}$.

Signed intersection over union is used in the ablation experiments of Sec. 5.5 of the main paper as an additional term in the payoff function $\pi(\hat{s}, s)$ as follows:

$$\pi(\hat{s}, s) = -\pi^*(\hat{s}, s) + \eta(s, \hat{s}) ,$$

$$\pi^*(\hat{s}, s) = \text{sIoU}(\boldsymbol{b}_s, \boldsymbol{b}_{\hat{s}}) + \|a_s^{y_s} - a_{\hat{s}}^{y_{\hat{s}}}\| + \frac{|t_s - t_{\hat{s}}|}{N_w} ,$$

where $\boldsymbol{b}_s$ denotes the bounding box of segment s.

References

[1] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *Journal on Image and Video Processing*, page 1, 2008. 1

[2] Roberto Henschel, Laura Leal-Taixe, Daniel Cremers, and Bodo Rosenhahn. Fusion of head and full-body detectors for multi-object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 2

[3] Margret Keuper, Siyu Tang, Bjorn Andres, Thomas Brox, and Bernt Schiele. Motion segmentation & multiple object tracking by correlation co-clustering. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 2

[4] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M Rehg. Multiple hypothesis tracking revisited. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 2

[5] 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. 2

[6] Andrea Simonelli, Samuel Rota Bulò, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[7] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2