

Supplementary Material for One Shot Deep Model for End-to-End Multi-Person Activity Recognition

Shuhei Tarashima
tarashima@acm.org

Innovation Center
NTT Communications Corp.
Tokyo, Japan

A Embedding Extraction Strategies *cf.* §2.1

In the multi-branch CNN of our TrAct-Net, embeddings with respect to re-ID and action/activity classification are extracted from the same location in the corresponding embedding maps with peaks of the detection heatmap. Here we compare this strategy to alternative sampling techniques. Specifically, following [8], we adopt RoI-Align applied in [8] and POS-Anchor applied in [9]. Notice that in all the cases we use the same architecture with TrAct-Net without embedding sampling modules in re-ID and action/activity branches. To make comparison simple, we only feed action/activity embeddings into the relation encoder in all the cases. The results are shown in Table 1. As we mentioned in §2.1 of the main paper, our approach achieves superior performance to these alternatives.

B Re-ID Loss Computation Strategies *cf.* §3.1.2

In the training of our TrAct-Net we compute the re-ID loss *sequence-wise* instead of *batch-wise*, since in the Volleyball dataset [9] instances are guaranteed to have the same identities *only* within the same sequence. To evaluate the influence of this trick, we use the Collective dataset [10] with a full annotation created by us. Since in our new annotation any identity of the same instance is guaranteed to be the same through the dataset, we can use classification loss [8] and *batch-wise* triplet loss in addition to proposed *sequence-wise* triplet loss (*cf.* §3.1.2 in the main paper) to define the re-ID loss. Table 2 presents the result of TrAct-Net with different re-ID loss functions. Interestingly, even when we only use sequence-wise supervision to compute re-ID loss, the performance is very competitive to alternatives which compute the loss through the batch. This indicates the effectiveness of our re-ID loss computation strategy under incomplete annotation for identity.

Table 1: Results of different embedding extraction strategies on the Volleyball dataset [1].

	Detection	Tracking		Action	Activity
	AP	IDF1	MOTA	mAP	Accuracy
RoI-Align [1]	93.3	95.5	91.3	44.8	93.7
POS-Anchor [1]	93.4	95.4	91.1	45.1	93.9
Ours	93.3	95.7	91.4	45.7	94.5

Table 2: Results of different re-ID loss computation strategies on the Collective [1] dataset.

	Detection	Tracking		Action	Activity
	AP	IDF1	MOTA	mAP	Accuracy
Batch-wise ([1])	98.6	91.4	81.3	47.1	92.1
Batch-wise (TrAct-Net)	98.5	91.4	81.1	46.9	92.1
Sequence-wise (TrAct-Net)	98.6	91.3	81.1	47.0	91.9

References

- [1] W. Choi, K. Shahid, and S. Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *ICCV Workshops*, 2009.
- [2] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, 2016.
- [3] P. Voigtlaender, M. Krause, A. Ösep, and J. Luiten. Mots: Multi-object tracking and segmentation. In *CVPR*, 2019.
- [4] Z. Wang, L. Zheng, Y. Liu, and S. Wang. Towards real-time multi-object tracking. *arXiv preprint arxiv:1909.12605*, 2019.
- [5] Y. Zhang, C. Wang, X. Wangy, W. Zeng, and W. Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *arXiv preprint arXiv:2004.01888*, 2020.