SUPPLEMENTARY MATERIAL FOR A CONFIDENCE-AWARE MATCHING STRATEGY FOR GENERALIZED MULTI-OBJECT TRACKING

Kyujin Shim, Jubi Hwang, Kangwook Ko, Changick Kim

Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Republic of Korea

1. TRAINING DETAILS

The detector models employed for the MOT17 [1] and MOT20 [2] test sets are obtained from ByteTrack [3]. Specifically, the model for the MOT17 test set is trained using a combination of the CrowdHuman [4], Cityperson [5], and ETHZ [6] datasets. Similarly, the detector for the MOT20 test set is trained exclusively with the CrowdHuman dataset. For DanceTrack [7], the detector is trained using 8 RTX 3090 GPUs, following the identical configuration outlined in ByteTrack, with the addition of the CrowdHuman dataset. Lastly, the detector models for MOT17-val and MOT20-val are acquired from ByteTrack and SparseTrack [8], respectively, with each model trained using its corresponding halved training set and the CrowdHuman dataset. The feature extractor models for MOT17 and MOT20 are obtained from BoT-SORT [9], while the extractor model for DanceTrack is acquired from Deep OC-SORT [10]. Each model is configured as an SBS50 architecture within the FastReID [11] framework and trained on its respective dataset using the default training settings provided by the framework.

2. DATASET DETAILS

The MOT17 [1], MOT20 [2], and DanceTrack [7] datasets used in this study contain 5,316, 8,931, and 41,796 training images, respectively. Also, 19,370, 2,500, and 2,056 person detection images from the CrowdHuman [4], Cityperson [5], and ETHZ [6] datasets, respectively, supplemented the training data. The average number of objects per image for training sets of MOT17, MOT20, DanceTrack, CrowdHuman, Cityperson, and ETHZ are approximately 21.1, 127.0, 8.4, 6.6, and 8.2, respectively.

3. COMPUTATIONAL COST

Table 1 shows the computational costs of each component in our confidence-aware matching strategy. Each only slightly increases response time, from 0.00352 to 0.00379 seconds for the MOT17 validation set and 0.00164 to 0.00173 seconds for the DanceTrack validation set. As a result, our tracker requires only 0.00379 and 0.00173 seconds per image for MOT17 and DanceTrack validation sets, respectively. Note that every FPS and time is measured only for the tracking algorithm without object detection and feature extraction.

Components			**MOT17-val**		**Dance-val**	
CCM	CFU	CMF	FPS	Time (s)	FPS	Time (s)
✗	✗	✗	283.8	0.00352	609.9	0.00164
✓	✗	✗	269.6	0.00371	592.0	0.00169
✓	✓	✗	264.4	0.00378	579.5	0.00173
✓	✓	✓	263.5	0.00379	577.8	0.00173

Table 1. The computational overheads of each proposed component in our confidence-aware matching strategy. CCM, CFU, and CMF denote confidence-aware cascade matching, confidence-aware feature update, and confidence-aware metric fusion, respectively. FPS represents frame per second, and time represents the tracking time required for each frame, measured in seconds.

4. ABLATIVE STUDY

In Table 2, we can confirm the impact of the β in our confidence-aware feature update (CFU). The case of $\beta = 0.95$ shows the most balanced results with the highest HOTA, IDF1, and AssA scores on both the MOT17 and DanceTrack validation sets. Therefore, we selected β to 0.95 as the default configuration. Note that setting the β as 0.95 performs adequately across all datasets, which contain diverse tracking scenes, and is expected to be a general real-world option. These results indicate that our method does not require complicated adjustments of the β for each tracking scene.

5. REFERENCES

[1] Anton Milan, Laura Leal-Taixe, Ian Reid, Stefan Roth, and Konrad Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv:1603.00831*, 2016.

	MOT17-val				**MOT20-val**				**DanceTrack-val**			
β	HOTA↑	IDF1↑	MOTA↑	AssA↑	HOTA↑	IDF1↑	MOTA↑	AssA↑	HOTA↑	IDF1↑	MOTA↑	AssA↑
0.85	70.2	83.1	**80.2**	72.7	**70.2**	**84.9**	**88.9**	**67.6**	60.8	62.3	**90.9**	46.3
0.90	70.1	83.0	80.0	72.7	70.0	84.6	**88.9**	67.3	60.8	62.5	**90.9**	46.4
0.95	**70.3**	**83.4**	79.6	**73.3**	69.7	84.2	88.8	66.8	**61.4**	**62.7**	**90.9**	**47.3**

Table 2. An ablation study for the β of our confidence-aware feature update (CFU) with the MOT17, MOT20, and DanceTrack validation sets.

[2] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé, “Mot20: A benchmark for multi-object tracking in crowded scenes,” *arXiv:2003.09003*, 2020.

[3] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang, “Bytetrack: Multi-object tracking by associating every detection box,” in *ECCV*, 2022.

[4] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun, “Crowdhuman: A benchmark for detecting human in a crowd,” *arXiv preprint*, vol. arXiv:1805.00123, 2018.

[5] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele, “Citypersons: A diverse dataset for pedestrian detection,” in *CVPR*, 2017.

[6] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool, “A mobile vision system for robust multi-person tracking,” in *CVPR*, 2008.

[7] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo, “Dancetrack: Multi-object tracking in uniform appearance and diverse motion,” in *CVPR*, 2022.

[8] Zelin Liu, Xinggang Wang, Cheng Wang, Wenyu Liu, and Xiang Bai, “Sparsetrack: Multi-object tracking by performing scene decomposition based on pseudo-depth,” *arXiv:2306.05238*, 2023.

[9] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky, “Bot-sort: Robust associations multi-pedestrian tracking,” *arXiv:2206.14651*, 2022.

[10] Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani, “Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification,” *arXiv:2302.11813*, 2023.

[11] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei, “Fastreid: A pytorch toolbox for general instance re-identification,” *arXiv:2006.02631*, 2020.