7. Supplementary

This supplementary material includes additional details regarding the definitions of the evaluation metrics for MTMC tracking as well as MTSC tracking, which are partially explained in Section 3.5. The measurements are adopted from the MOTChallenge [5, 24] and DukeMTMC [34] benchmarks. Besides, the performance of our baseline image-based ReID methods in terms of mAP measured by the top 100 matches for each query is presented, which is the metric used in our evaluation server.

7.1. Metrics in CLEAR MOT

The CLEAR MOT [5] metrics are used in the MOTChallenge benchmark for evaluating multiple object tracking performance. The distance measure, *i.e.*, how close a tracker hypothesis is to the actual target, is determined by the intersection over union between estimated bounding boxes and the ground truths. The similarity threshold for true positives is empirically set to 50%.

The Multiple Object Tracking Accuracy (MOTA) combines three sources of errors to evaluate a tracker's performance.

$$\text{MOTA} = 1 - \frac{\sum_t \left(\text{FN}_t + \text{FP}_t + \text{IDSW}_t\right)}{\sum_t \text{GT}_t}, \tag{1}$$

where t is the frame index. FN, FP, IDSW and GT respectively denote the numbers of false negatives, false positives, identity switches and ground truths. The range of MOTA in percentage is $(-\infty, 100]$, which becomes negative when the number of errors exceeds the ground truth objects.

Multiple Object Tracking Precision (MOTP) is used to measure misalignment between annotated and predicted object locations, defined as

$$\text{MOTP} = 1 - \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t}, \tag{2}$$

in which c_t denotes the number of matches and $d_{t,i}$ is the bounding box overlap between target i and the ground truth at frame index t. According to the analysis in [19], MOTP shows a remarkably low variation across different methods compared with MOTA. Therefore, MOTA is considered as a more reliable evaluation for tracking performance.

Besides MOTA and MOTP, there are other metrics for evaluating the tracking quality. MT measures the number of mostly tracked targets that are successfully tracked by at least 80% of their life span. On the other hand, ML calculates the number of mostly lost targets that are only recovered for less than 20% of their total lengths. All the other targets are classified as partially tracked (PT). Furthermore, FAR measures the average number of false alarms, *i.e.*, FN, FP and IDSW, per frame.

Norm	Rank-100 mAP
Bhattacharyya	**5.1%**
L_2	5.0%
L_1	4.8%
L_∞	2.5%

Table 10. Performance of non-metric learning methods using CNN features extracted from FVS [43] on our CityFlow-ReID benchmark, showing rank-100 mAP, corresponding to the experimental results of Tab. 3.

7.2. Metrics in DukeMTMC

There are three evaluation metrics introduced by the DukeMTMC benchmark, namely identification precision (IDP), identification recall (IDR), and the F1 score IDF1. They are defined based on the counts of false negative identities (IDFN), false positive identities (IDFP) and true positive identities (IDTP), which are defined as follows,

$$\text{IDFN} = \sum_{\tau \in \text{AT}} \sum_{t \in \mathcal{T}_\tau} m\left(\tau, \gamma_m\left(\tau\right), t\right), \tag{3}$$

$$\text{IDFP} = \sum_{\gamma \in \text{AC}} \sum_{t \in \mathcal{T}_\gamma} m\left(\tau_m\left(\gamma\right), \gamma, t\right), \tag{4}$$

$$\text{IDTP} = \sum_{\tau \in \text{AT}} \|\tau\| - \text{IDFN} = \sum_{\gamma \in \text{AC}} \|\gamma\| - \text{IDFP}, \tag{5}$$

where τ and γ respectively denotes the true and computed trajectories, AT and AC are all true and computed identities, and $\mathcal{T}$ represents the set of frame indices t over which the corresponding trajectory extends. $\|\cdot\|$ returns the number of detections in a given trajectory. The expression $m\left(\tau, \gamma, t\right)$ calculates the number of missed detections between τ and γ along time. We use $\gamma_m\left(\tau\right)$ and $\tau_m\left(\gamma\right)$ to denote the bipartite match from τ to γ and *vice versa*, respectively. Identification precision (recall) is defined as the ratio of computed (true) detections that are correctly identified.

$$\text{IDP} = \frac{\text{IDTP}}{\text{IDTP} + \text{IDFP}}, \tag{6}$$

$$\text{IDR} = \frac{\text{IDTP}}{\text{IDTP} + \text{IDFN}}. \tag{7}$$

IDF1 is the fraction of correctly identified detections over the average number of true and computed detections.

$$\text{IDF1} = \frac{2 \cdot \text{IDF1}}{2 \cdot \text{IDF1} + \text{IDFP} + \text{IDFN}}. \tag{8}$$

Compared to the metrics in CLEAR MOT, the truth-to-result mapping in IDF1 computation is not frame-by-frame but identity-by-identity for the entire sequence, and the errors of any type are penalized based on binary mismatch. Therefore, IDF1 can handle overlapping and disjoint fields of view for the evaluation of MTMC tracking performance, which is a property absent in all previous measures.

Loss	ResNet50 [12]	ResNet50M [54]	ResNeXt101 [51]	SEResNet50 [16]	SEResNeXt50 [16]	DenseNet121 [17]	InceptionResNetV2 [38]	MobileNetV2 [36]
Xent [40]	20.3%	20.4%	**21.6%**	18.6%	21.5%	18.6%	16.2%	**10.4%**
Htri [13]	22.1%	21.3%	23.3%	19.8%	21.7%	**24.0%**	17.8%	0.0%
Cent [48]	5.6%	6.1%	6.2%	8.3%	8.4%	**9.5%**	4.9%	5.2%
Xent+Htri	**23.7%**	**24.2%**	**26.3%**	**24.3%**	**25.1%**	**26.0%**	**20.5%**	6.5%
Xent+Cent	17.8%	21.7%	19.5%	20.9%	23.2%	**23.3%**	18.8%	8.3%

Table 11. Performance of state-of-the-art metric learning methods for person ReID on CityFlow-ReID, showing rank-100 mAP, corresponding to the experimental results of Tab. 4. The best architecture and loss function are highlighted for each row/column, respectively, with the shaded cells indicating the overall best.

Method	Rank-100 mAP
MobileNetV1+BA [18]	25.6%
MobileNetV1+BH [18]	**26.5%**
MobileNetV1+BS [18]	25.6%
MobileNetV1+BW [18]	25.4%

Table 12. Performance of the state-of-the-art metric learning method for vehicle ReID, with different sampling variants, on CityFlow-ReID, corresponding to the experimental results of Tab. 6. Rank-100 mAP is shown.

7.3. Rank-K mAP for evaluating image-based ReID

As mentioned in Section 3.5, to measure the total mAP of each submission, a distance matrix of dimension $Q \times T$ is required, where Q and T are the numbers of queries and test images, respectively. For an evaluation server with many users and each of them is allowed to submit multiple times, such large file size may lead to system instability. Thus, we create a new evaluation metric for image-based ReID, named rank-K mAP, that measures the mean of average precision for each query considering only the top K matches, so that the required dimension of each submission can be reduced to $Q \times K$. Note that K usually needs to be larger than the maximum length of ground-truth trajectories, which is chosen to be 100 for our evaluation.

Because rank-100 mAP is adopted in our evaluation server, we present here the additional experimental results in Tab. 10, Tab. 11 and Tab. 12, which correspond to Tab. 3, Tab. 4 and Tab. 6, respectively.