A. Additional Experimental Results

A.1. Paths for Evaluation

In our paper, the instance path is only applied in training, facilitating the identity feature learning of the main search path. During inference, we drop the instance path and the images are only passed through the search path. We compare the results of using different paths for testing, as shown in Tab. 5. It can be seen that using two paths for evaluation cannot bring extra performance gains. This indicates the context-invariant embeddings produced by our framework.

Table 5. Comparisons of using different paths for evaluation on the CUHK-SYSU dataset.

Inference path	mAP	Rank-1
Two paths	85.65	86.73
Search path	**85.72**	**86.86**

A.2. Different Detection Networks

Following [16], we choose the RepPoints as the detection network. To show the expandability of our R-SiamNet, different detectors are incorporated into our framework, including Faster R-CNN [28], RetinaNet [23] and RepPoints [38]. As reported in Tab. 6, the final performance gaps among different detectors are small, exhibiting the effectiveness and robustness of our framework.

Table 6. Comparisons when incorporated with different detectors on the CUHK-SYSU dataset.

Detector	mAP	Rank-1
Faster R-CNN	84.84	85.72
RetinaNet	85.39	86.59
RepPoints	**85.72**	**86.86**

A.3. Comparisons with Two-Step Manner

We combine a well-trained RepPoints detector [38] and an unsupervised re-ID model called SPCL [13] as our two-step competitor. As shown in Tab. 7, our method outperforms it by a large margin with higher efficiency. It shows that training detection and identification end-to-end is beneficial for obtaining better representations. It may also imply the importance of instance-level consistency learning.

Table 7. Comparisons with two-step manner on the CUHK-SYSU dataset.

Methods	mAP	Rank-1
RepPoints+SPCL	73.43	74.79
Ours	**85.72**	**86.86**

A.4. Evaluation on filter strategy in the clustering.

To analyze the effectiveness of the filter strategy in clustering, we conduct experiments with/without filtering by image information. As shown in Tab. 8, we observe 1.22%/1.13% rank-1 drops on CUHK-SYSU/PRW datasets by removing the filter strategy. This shows that it is beneficial to filter the aggregation of the persons from the same scene images.

Table 8. Performance of our method with/without the filter strategy in clustering. Results on the CUHK-SYSU and PRW datasets are shown. R-SiamNet w/o filter means the clustering is applied without filtering by image information.

Methods	CUHK-SYSU		PRW	
	mAP	Rank-1	mAP	Rank-1
R-SiamNet w/o filter	84.74	85.64	20.31	72.23
R-SiamNet	**85.72**	**86.86**	**21.16**	**73.36**

A.5. Different numbers of training epochs

We illustrate the mAP scores with different numbers of training epochs. As Fig. 5 shows, the results of three data scales are exhibited on the CUHK-SYSU dataset. It can be observed that the performance improves steadily to saturation as the epoch increases. With smaller data scales, the mAP reaches saturation earlier.

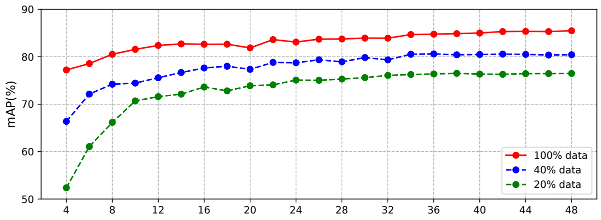
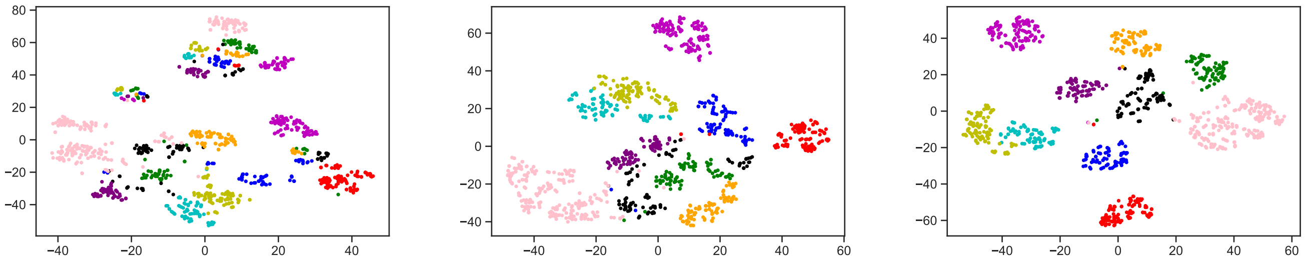


Figure 5. Performance with different numbers of training epochs. The results of three data scales are exhibited on the CUHK-SYSU dataset.

A.6. Runtime Comparisons

To compare the evaluation efficiency of our framework with other methods, we report the average runtime of the inference stage for a whole scene image, shown in Tab. 9. Since these methods are evaluated with different GPUs, we also exhibit the Tera-Floating Point Operation per-second (TFLOPs) for fair comparisons. Similar to other methods [5, 26, 4], we evaluate the models with an input image size of 900 × 1500. As shown in Tab. 9, our R-SiamNet takes 72 milliseconds to process one image, which is faster than the two-step method MGTS [4] by a large margin. The query-guided method QEEPS [26] requires to re-compute



(a) R-SiamNet w/o L_{ins} & L_{int} & L_{clu}

(b) R-SiamNet w/o L_{clu}

(c) R-SiamNet

Figure 6. T-SNE feature visualization on a part of the PRW training set (10 classes, $1,089$ pedestrians). (a) R-SiamNet without L_{ins} & L_{int} & L_{clu}. (b) R-SiamNet without L_{clu}. (c) Our proposed R-SiamNet with both instance-level consistency learning and cluster-level contrastive learning. Colors denote person identities.

all the gallery embeddings for each query image. This time-consuming operation reduces the evaluation efficiency. Moreover, our method is 13% faster than NAE [5]. These results clearly demonstrate the efficiency of our R-SiamNet in evaluation.

Table 9. Runtime comparisons of different methods when evaluation. The average runtime for one image with the size of 900×1500 is exhibited on the CUHK-SYSU dataset.

Methods	GPU (TFLOPs)	Runtime (ms)
MGTS [4]	K80 (8.7)	1269
QEEPS [26]	P6000 (12.0)	300
NAE+ [5]	V100 (14.1)	98
NAE [5]	V100 (14.1)	83
Ours	V100 (14.1)	**72**

B. More Qualitative Analysis

B.1. Feature Visualization

To analyze the discriminative ability of our learned features, we employ the t-SNE [32] to visualize the feature representations in training. As illustrated in Fig. 6, there are $1,089$ pedestrians with 10 classes, which is a subset of the PRW training set. Different colors represent different classes.

Fig. 6(a) shows the result when training with a single search path. Without applying the instance-level consistency learning and cluster-level contrastive learning, the learned features show large intra-class distances and small inter-class distances. When adding the instance-level consistency learning, including L_{ins} and L_{int}, the result is shown in Fig. 6(b). It is observed that the feature embeddings of the same category can be aggregated compared with Fig. 6(a). This shows the effectiveness of our instance-level consistency learning. Nevertheless, the features within the class are gathered loosely, and the margins among different classes are not clear. Furthermore, we apply the cluster-level contrastive learning, and the result is shown in Fig. 6(c). It can be seen that both intra-class compactness and inter-class separability are further encouraged. There are obvious margins among most categories. This verifies that our R-SiamNet can generate discriminative embeddings under the weakly supervised settings.