Supplementary Materials of Optimal Transport for Label-Efficient Visible-Infrared Person Re-Identification

Jiangming Wang[1], Zhizhong Zhang[1(✉)], Mingang Chen[2], Yi Zhang[3], Cong Wang[4], Bin Sheng[5], Yanyun Qu[6], and Yuan Xie[1(✉)]

[1] East China Normal University, Shanghai, China
[2] Shanghai Development Center of Computer Software Technology, Shanghai, China
[3] ZheJiang Lab, Hangzhou, China
[4] Huawei Technologies, Hangzhou, China
[5] Shanghai Jiao Tong University, Shanghai, China
[6] Xiamen University, Fujian, China
{51215901073}@stu.ecnu.edu.cn, {zzzhang,yxie}@cs.ecnu.edu.cn, {cmg}@sscenter.sh.cn, {zhangyi620}@zhejianglab.com, {wangcong64}@huawei.com, {shengbin}@sjtu.edu.cn, {yyqu}@xmu.edu.cnn

1 Introduction

In this supplementary material, we firstly illustrate the detail of network architectures and implementation in Sec. 2. Then, we provide additional infrared label distribution experimental results in Sec. 3. Besides, we will further discuss the effects of RGB source domain in Sec. 4. Finally, we would like to discuss the influence on performances by using different kinds of UDA-ReID or USL-ReID to generate visible pseudo labels.

2 Architectures and Implementation

In this section, we illustrate the detailed network architectures especially the modality classifier. Besides, we will provide some details of implementation.

2.1 Detailed Architectures of Network

We use the ResNet-50 [5] pre-training on the ImageNet [1] as our feature extracting backbone. While, modality classifier consists of three linear layers and one BN layer [6]. Detailed network structure is shown in Fig. 1. We concatenate the avgpooling features of all residual blocks of ResNet-50. Then, we feed them into the modality classifier module to calculate the discriminative loss ($\mathcal{L}_D$). More specifically, each avgpooling feature of residual block will go through an unique BN layer before concatenation.

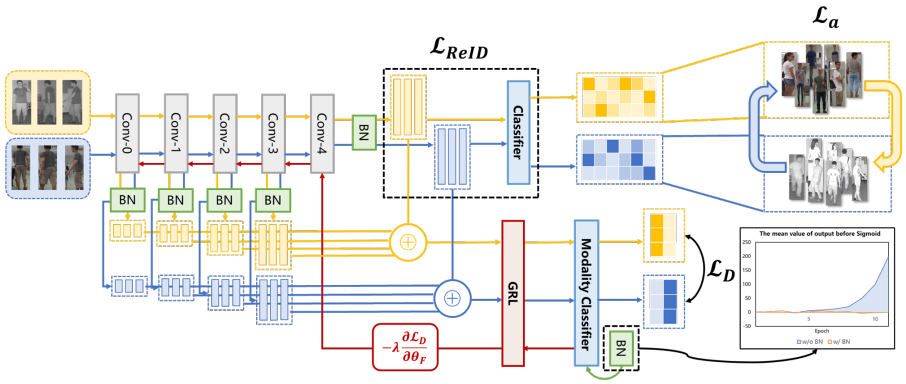


Fig. 1. Details of our proposed network structure. ⊕ means concatenation at the feature dimension.

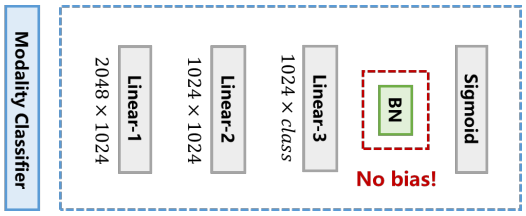


Fig. 2. Details of modality classifier structure. *class* means the number of classes (identities) for each dataset. The BN layer before the sigmoid function is deployed without bias.

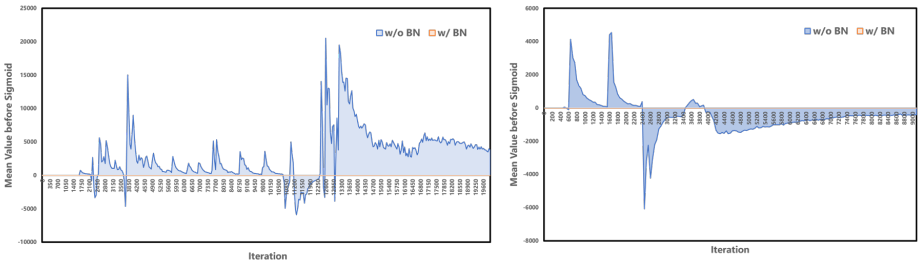


Fig. 3. Discussion of the BN layer for modality classifier. The X-axis indicates the iteration and the Y-axis indicates the mean value of output before sigmoid function. w/o BN means without BN layer and w/ BN means with BN layer.

2.2 Detailed Architectures of Modality Classifier.

As shown in Fig. 2, modality classifier is combined with three linear layers and a unique BN layer before sigmoid function. The hidden dimensions of the

first two linear layers are both 1024. The last linear layer is a classification layer and its dimension is equal to the number of classes (identities) for each dataset.

BN Layer of Modality Classifier. In addition, the BN layer of modality classifier is deployed without bias, which is important for the stability of training process. Fig. 3 illustrates that the output before sigmoid of modality classifier is not approximately zero-centered without BN layer. This phenomenon will lead to the value of $\mathcal{L}_D$ (binary loss) become very large because the output of sigmoid function is approximate to 1 (if output before sigmoid of modality classifier has greater positive value) or is approximate to 0 (if output before sigmoid of modality classifier has greater negative value). That means the other training losses cannot be completely optimized, which damages the performance seriously.

2.3 Detailed Implementation

GRL. The GRL layer has no parameters and γ is formulated as follows:

$$\gamma = \frac{2}{1 + \exp(-\tau \frac{iter}{maxiter})} - 1, \tag{1}$$

where τ is equal to 10 and $maxiter$ is equal to 10000. The $iter$ linearly increases as the training goes on.

OTLA. We need to offset the repeated infrared samples of each epoch before implementing OTLA because each unique infrared sample only can be assigned with one label. More specifically, we average the same samples' softmax output of classifier:

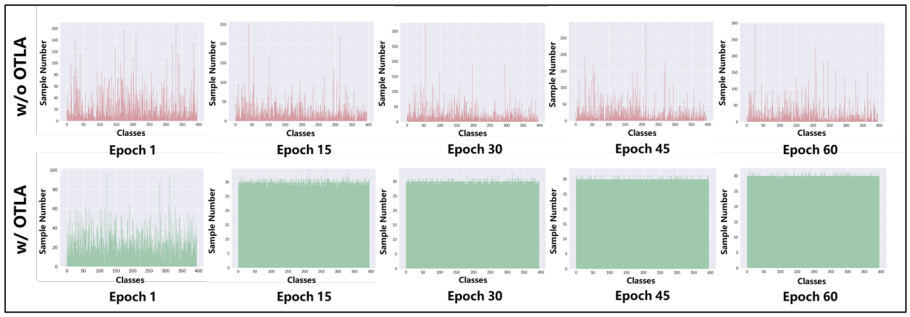
$$\boldsymbol{P}^r := \left\{ \boldsymbol{P}_k^r | \boldsymbol{P}_k^r = \frac{\boldsymbol{P}_i^r + \sum_{\boldsymbol{x}_i = \boldsymbol{x}_j} \boldsymbol{P}_j^r}{1 + \sum_{\boldsymbol{x}_i = \boldsymbol{x}_j} 1} \right\}, \tag{2}$$

where $\boldsymbol{x}_i = \boldsymbol{x}_j$ means $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are the same sample. Note Eq. (2) is only conducted to the repeated sample in $\{\boldsymbol{x}_i^r\}_{i=1}^m$, and the subscript k represents the row dimension. Besides, due to the unstable pseudo labels of infrared samples during training process, we multiply coefficients 0.1 and 0.5 for infrared modal cross-entropy and triplet loss, respectively.

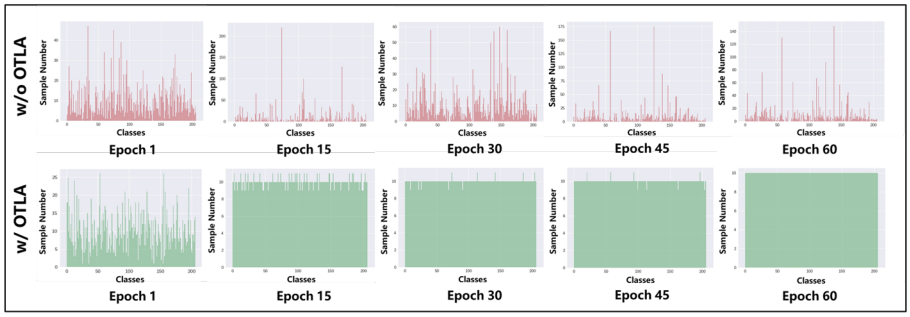
Visible Pseudo Label. When we obtain the visible pseudo labels from SpCL [4] (HCD [8] and MMT [3] as well), we filter visible samples labelled as outliers (pseudo labels produced by DBSCAN [2]) and then use remaining visible samples for following training.

3 Label Distribution of Infrared Images

As shown in Fig. 4, we can also find that OTLA alleviates the degeneration of the classifier on RegDB as training goes on.



(a) Infrared pseudo label distribution of SYSU-MM01



(b) Infrared pseudo label distribution of RegDB

Fig. 4. The changes of label distribution during the training process (green area means pseudo label distribution w/ OTLA, red area means pseudo label distribution w/o OTLA).

4 Effects of RGB Source Domain

In this section, we firstly show the performances on SYSU-MM01 and RegDB by using SpCL [4]. Then, we would like to visualize the quality of generated visible pseudo labels on SYSU-MM01.

4.1 Performance on USVI-ReID Setting

As shown in Tab. 1 and Tab. 2, we can find that adding other labeled visible datasets can achieve better performance on both SYSU-MM01 and RegDB. Besides, it seems that the size of extra labeled visible dataset achieve a little impact on performance. The possible reason may be the clustering results (the quality of pseudo labels) produced by SpCL [4] are close, which will be visualized in the next subsection.

4.2 Visualization of Visible Pseudo Label

In this subsection, we firstly define four metrics to measure the quality of visible pseudo labels produced by SpCL [4]. Then, we combine the above metrics

Table 1. Performance on USVI-ReID Setting of SYSU-MM01 by using SpCL [4].

Order	Dataset				All Search				Indoor Search			
	Unlabeled	Labeled										
	SYSU	Market	Duke	MSMT	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
1	✓	-	-	-	26.24	67.63	80.75	24.05	24.86	71.15	84.38	35.14
2	✓	✓	-	-	29.98	71.79	83.85	27.13	29.80	74.14	87.64	38.79
3	✓	-	✓	-	28.45	68.76	81.59	25.32	30.21	72.46	86.23	38.91
4	✓	-	-	✓	30.06	68.68	82.15	25.48	26.04	71.29	85.78	35.85

Table 2. Performance on USVI-ReID Setting of RegDB by using SpCL [4]

Order	Dataset				Visible2Thermal				Thermal2Visible			
	Unlabeled	Labeled										
	RegDB	Market	Duke	MSMT	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
1	✓	-	-	-	25.78	46.26	55.34	24.93	30.49	47.52	56.12	26.49
2	✓	✓	-	-	29.66	50.68	60.19	26.67	32.28	53.25	61.60	26.72
3	✓	-	✓	-	29.19	51.60	61.87	27.23	31.91	53.23	62.41	27.44
4	✓	-	-	✓	32.86	54.17	62.48	29.69	32.14	56.07	64.95	28.64

to form a complete one named R_{plq}. Notice that these metrics are only used for evaluation not for training. Besides, we will provide t-SNE [7] results to further visualize clustering process of unlabeled visible data on SYSU-MM01. Finally, we will display some concrete clustering results by using SpCL.

The Average Maximum Proportion of Ground-truth Classes (R_{gt}). For samples with ground-truth class i, we calculate the maximum pseudo class proportion R^i_{gt}, which can be formulated as:

$$R^i_{gt} = \frac{\max\limits_j |\boldsymbol{x}^{v_i}_{p_j}|}{|\boldsymbol{x}^{v_i}|} \in [0, 1], \tag{3}$$

where $|\boldsymbol{x}^{v_i}|$ means that the number of visible samples belong to ground-truth class i, $|\boldsymbol{x}^{v_i}_{p_j}|$ means that the number of visible samples with pseudo class j but belong to ground-truth class i. Then, we average R^i_{gt} for each ground-truth class to obtain the R_{gt}:

$$R_{gt} = \frac{1}{C_v} \sum_{i=1}^{C_v} R^i_{gt} \in [0, 1], \tag{4}$$

where C_v is the number of ground-truth visible classes. We can find that the higher R_{gt} indicates the samples with more consistent pseudo labels of each ground-truth class.

The Average Maximum Proportion of Pseudo Classes (R_{ct}). For samples with pseudo class j, we calculate the maximum ground-truth class proportion R^j_{ct}, which can be formulated as:

$$R^j_{ct} = \frac{\max\limits_i |\boldsymbol{x}^{p_j}_{v_i}|}{|\boldsymbol{x}^{p_j}|} \in [0, 1], \tag{5}$$

where $|\boldsymbol{x}^{p_j}|$ means that the number of visible samples belong to pseudo class j, $|\boldsymbol{x}_{v_i}^{p_j}|$ means that the number of visible samples with ground-truth class i but belong to pseudo class j. Then, we average R_{ct}^{j} for each pseudo class to obtain the R_{ct}:

$$R_{ct} = \frac{1}{C_v^{'}} \sum_{j=1}^{C_v^{'}} R_{ct}^{j} \in [0, 1], \tag{6}$$

where $C_v^{'}$ is the number of pseudo visible classes. Apparently, the higher R_{ct} indicates samples with more consistent ground-truth labels of each pseudo class.

The R_{gt} and R_{ct} are complementary with each other. As for R_{gt}, there is a special case that pseudo labels of all samples are the same, which lead to R_{gt} become to 1. Obviously, the quality of pseudo labels are much worse in that case. Fortunately, if R_{ct} is also higher, this case can possibly not happen. As for R_{ct}, there also exists a bad case that samples of one ground-truth class with different classes of pseudo labels, which lead to high value of R_{ct}. However, higher R_{gt} can avoid this case.

The Proportion of Visible Training Samples (P_v). Clustering methods (*e.g.*, DBSCAN [2]) will produce outlier samples, which do not belong to any existing pseudo class. Therefore, we need to filter them. In this case, the number of visible training samples will be reduced, which may be harmful to the following training process. Considering that, we define the P_v, the proportion of visible training samples, as another metric:

$$P_v = \frac{N_v^{'}}{N_v} \in [0, 1], \tag{7}$$

where N_v is the number of all visible samples and $N_v^{'}$ is the number of visible samples after filtering. If filtering step is not involved, the P_v is equal to 1.

The Credibility of Visible Pseudo Classes (Q_v). If the number of pseudo classes $C_v^{'}$ is close to the number of ground-truth classes C_v, the quality of pseudo labels is possibly higher. Therefore, the Q_v is formulated as:

$$Q_v = \frac{\min\{C_v, C_v^{'}\}}{\max\{C_v, C_v^{'}\}} \in [0, 1], \tag{8}$$

the only choice for Q_v to become 1 is that $C_v^{'}$ is equal to C_v.

The Final Metric R_{plq}. The final metric is designed as follows:

$$R_{plq} = \frac{R_{gt} + R_{ct}}{2} \cdot P_v \cdot Q_v \in [0, 1], \tag{9}$$

the higher R_{plq} means higher quality of pseudo labels.

After defining the above metrics, we can visualize the pseudo labels' quality produced by SpCL [4]. As shown in Fig. 5, we can find that values of all metrics are near for each setting, which indicates the quality of pseudo labels produced by SpCL is close.

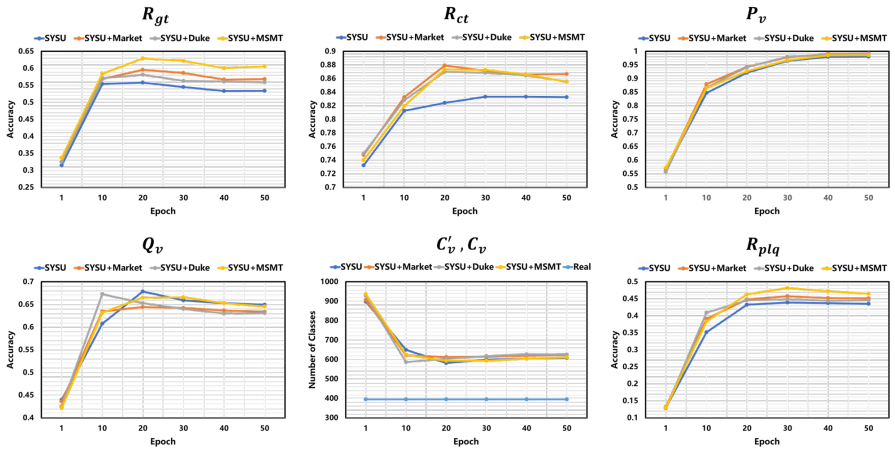


Fig. 5. The quality of pseudo labels produced by SpCL [4] on SYSU-MM01. In addition to the mentioned five metrics (including final metric R_{plq}), we also visualize the change of the number of pseudo classes ($C_v^{'}$) during the training process. *Real* means the number of ground-truth classes (C_v).

2D Projection of the Embedding Space by Using t-SNE. As shown in the Fig. 6, we visualize the clustering process of SpCL [4] every 10 training epochs (different colors means different kinds of pseudo classes and each figure contains visible samples of 20 real classes). With or without other annotated datasets, the clustering results are very similar, which can further illustrate that the quality of pseudo labels is close. Besides, the clustering pseudo classes is much more than real classes.

Concrete Clustering Results of SpCL. As shown in the Fig. 7, we display the images with the same pseudo class when we complete clustering on 'SYSU' setting (only involve visible data of unlabeled SYSU-MM01) by using SpCL. It appears that SpCL allows us to find appearance similar samples.

Table 3. Performance on USVI-ReID Setting of SYSU-MM01 by using HCD [8].

Order	Dataset				All Search				Indoor Search			
	Unlabeled	Labeled										
	SYSU	Market	Duke	MSMT	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
1	✓	-	-	-	17.14	53.88	69.52	16.76	19.74	63.90	80.71	28.57
2	✓	✓	-	-	17.41	54.04	69.71	16.71	22.01	64.99	83.06	30.57
3	✓	-	✓	-	17.04	55.32	71.02	17.03	20.20	62.14	79.89	28.83
4	✓	-	-	✓	17.62	55.04	70.10	16.82	20.24	63.22	79.26	29.61

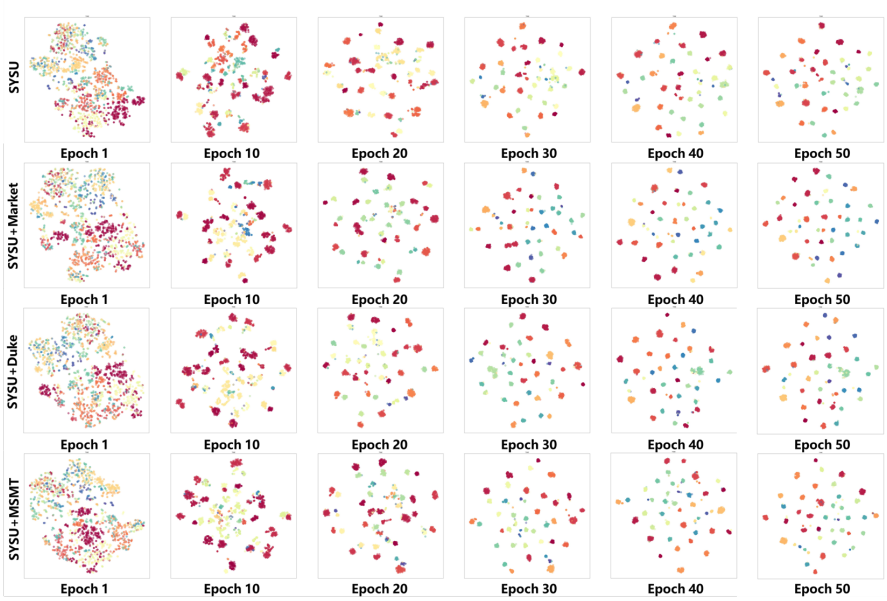


Fig. 6. 2D projection of the embedding space on SYSU-MM01 produced by SpCL [4] using t-SNE. Different colors means different kinds of pseudo classes.

Table 4. Performance on USVI-ReID Setting of SYSU-MM01 by using MMT† [3]. † indicates that we use DBSCAN version of MMT.

Order	Dataset				All Search				Indoor Search			
	Unlabeled	Labeled										
	SYSU	Market	Duke	MSMT	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
1	✓	✓	-	-	23.46	63.95	78.20	20.35	21.15	63.04	80.12	29.81
2	✓	-	✓	-	25.24	65.66	79.57	22.08	23.73	66.58	81.16	32.25
3	✓	-	-	✓	23.90	62.66	77.47	21.49	22.24	64.27	79.89	30.67

5 Effects of Different UDA-ReID or USL-ReID Methods

In this section, we would like to discuss the effects of our proposed pipeline performances on SYSU-MM01 dataset by using different UDA-ReID or USL-ReID methods. For more specific, we adopt the latest one-stage method HCD [8] and two-stage method MMT(DBSCAN version) [3] to generate visible pseudo labels. Besides, we find that our proposed pipeline can achieve better performance if generated visible pseudo labels obtain higher R_{plq}, which indicates that R_{plq} may be positively correlated with the quality of generated visible pseudo labels.

5.1 Performance on USVI-ReID Setting of SYSU-MM01

As shown in Tab. 3 and Tab. 4, we surprisingly discover that the performance impact brought by source domain is less significant. However, performance

Fig. 7. Concrete clustering results of 'SYSU' setting (only involve visible data of unlabeled SYSU-MM01) by using SpCL [4]. Each row shows the images assigned to a specific pseudo class. 'ID' represents their ground-truth label.

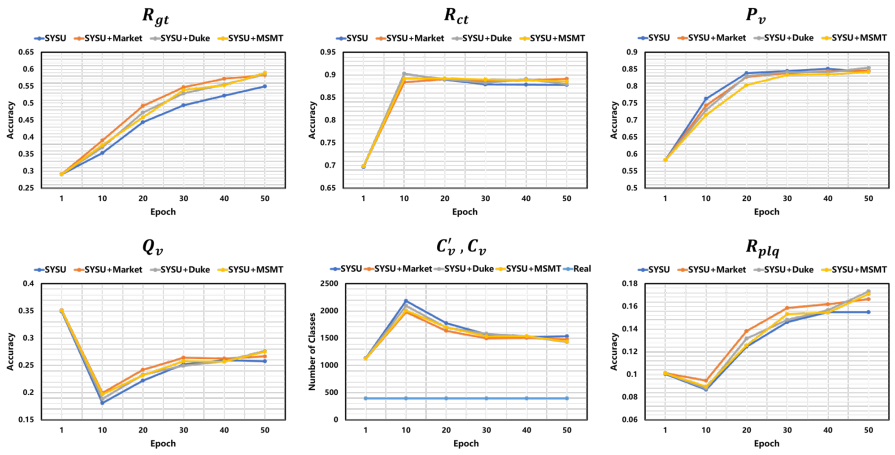


Fig. 8. The quality of pseudo labels produced by HCD [8] on SYSU-MM01.

changes significantly when employ different UDA-ReID or USL-ReID methods, which indicates that the quality of visible pseudo labels varies differently. With the performance of Tab. 1, we can conclude that SpCL [4] can produce visible pseudo labels of possibly high quality though there's still a certain gap from real labels.

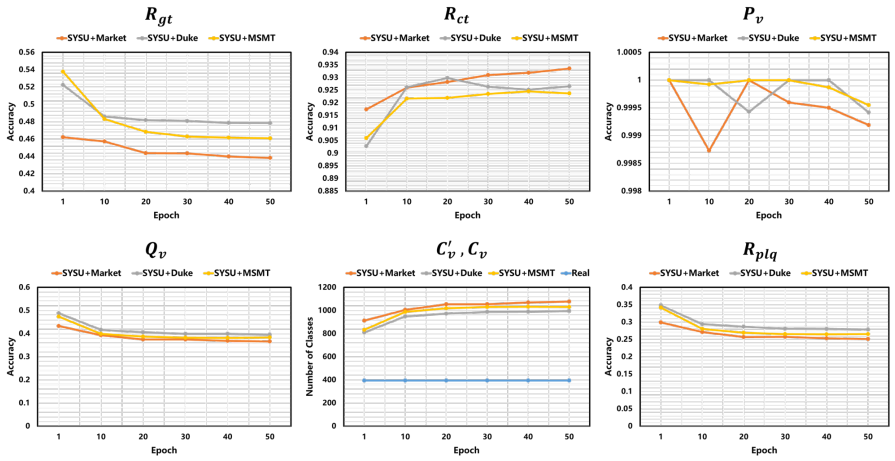


Fig. 9. The quality of pseudo labels produced by MMT (DBSCAN version) [3] on SYSU-MM01.

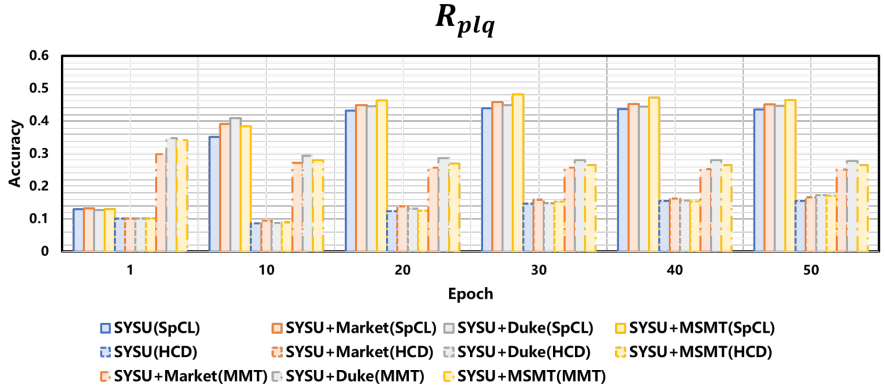


Fig. 10. The value of R_{plq} with different settings (*i.e.,* different UDA-ReID or USL-ReID methods, with or without extra annotated visible datasets) on SYSU-MM01.

5.2 The Relation between Performance and R_{plq}

Considering the performances of Tab. 1 and Tab. 3 and Tab. 4, we get all values of R_{plq} together as shown in the Fig. 10. Then, we can find that higher R_{plq} means better performance, which indicates that R_{plq} may be positively correlated with the quality of generated visible pseudo labels.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. IEEE (2009) 1

2. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: TKDD. vol. 96, pp. 226–231 (1996) 3, 6
3. Ge, Y., Chen, D., Li, H.: Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. ICLR (2020) 3, 8, 10
4. Ge, Y., Zhu, F., Chen, D., Zhao, R., Li, H.: Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. NIPS (2020) 3, 4, 5, 6, 7, 8, 9
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 1
6. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. pp. 448–456. PMLR (2015) 1
7. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. JMLR **9**(11) (2008) 5
8. Zheng, Y., Tang, S., Teng, G., Ge, Y., Liu, K., Qin, J., Qi, D., Chen, D.: Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In: CVPR. pp. 8371–8381 (2021) 3, 7, 8, 9