Optimal Transport for Label-Efficient Visible-Infrared Person Re-Identification

Jiangming Wang, Zhizhong Zhang(✉), Mingang Chen, Yi Zhang, Cong Wang,
Bin Sheng, Yanyun Qu, and Yuan Xie(✉)

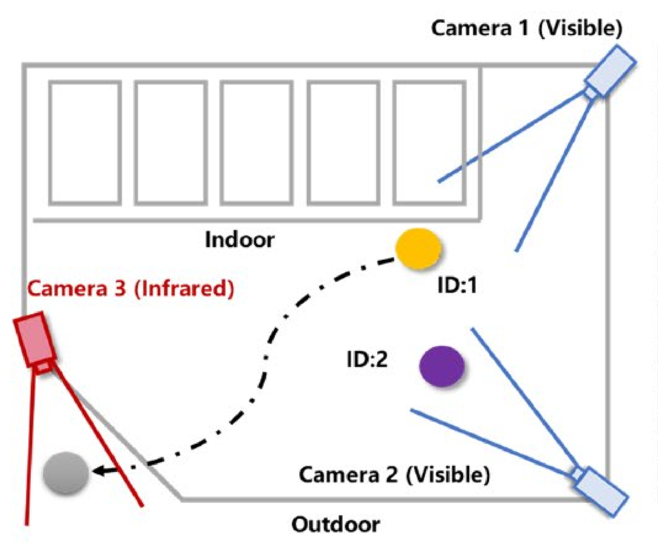
I. Introduction

Observation:

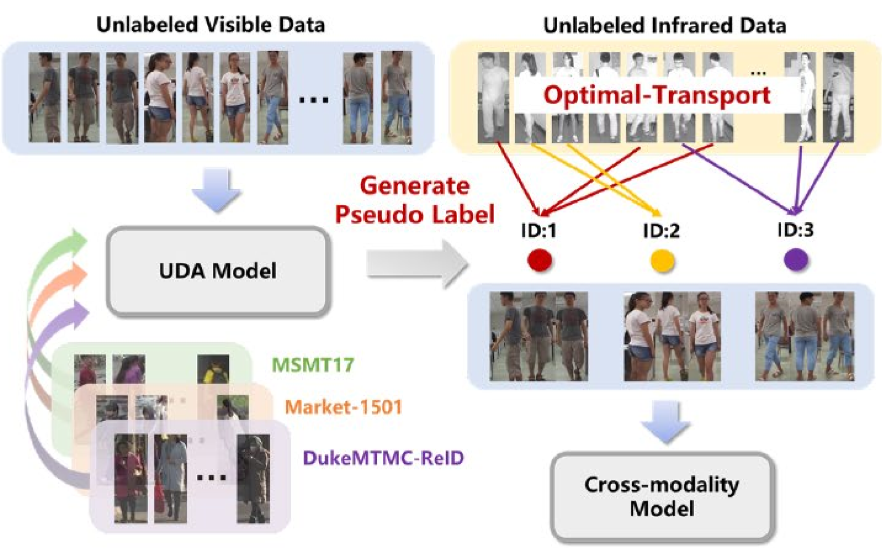
- The scale of VI-ReID datasets are relatively small
- Infrared data without color information is hard to label
- Expensive laboring efforts of annotation
- Visible ReID datasets have rich annotation information
- SOTA clustering-based UDA-ReID methods can generate reliable pseudo labels for popular visible ReID datasets

Question:

- **Can we learn a VI-ReID model only with visible modal supervision or even without supervision ?**
- **Can we get help of other rich annotated visible ReID datasets ?**



Real-Life Scenario for Pipeline Application



Label-Efficient Learning Pipeline for VI-ReID

II. Methodology

The First Stage (If unsupervised):

- Using clustering-based UDA-ReID method to generate reliable visible pseudo labels. In our paper, we adopt SpCL (Ge *et al.* NIPS, 2020) as pseudo label generator.

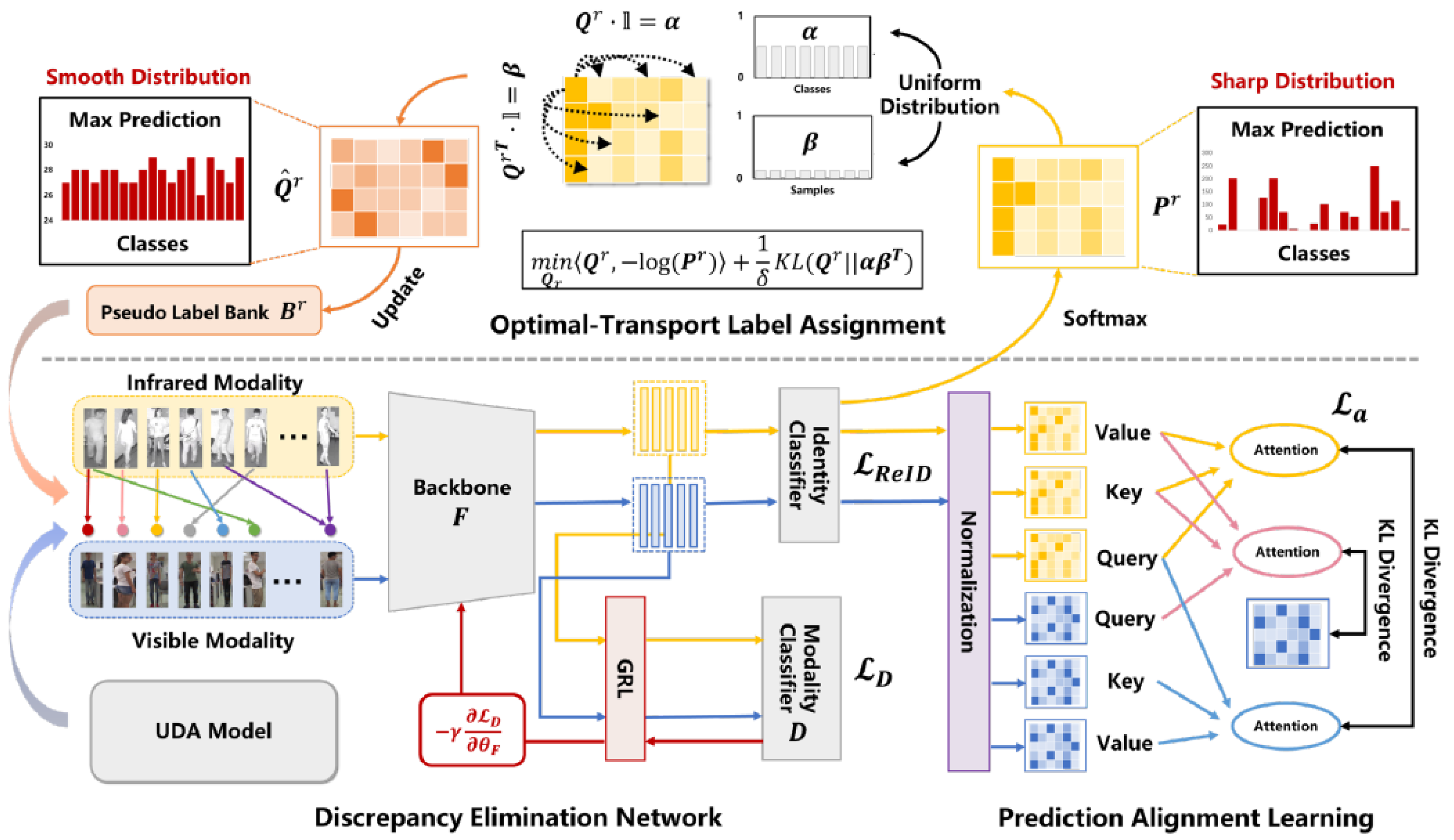
The Second Stage:

Discrepancy Elimination Network

- A backbone network aims to reduce the modality gap. Let f_i^v, f_i^r denote the feature output of backbone $\boldsymbol{F}$ and $\boldsymbol{D}$ denote modality classifier:

$$\mathcal{L}_{ReID} = \mathcal{L}_{CE} + \mathcal{L}_{Tri},$$

$$\mathcal{L}_D = \max_{\boldsymbol{F}}\min_{\boldsymbol{D}} \mathbb{E}_{f_i^v}\left[\log\left(1 - D(f_i^v)\right)\right] + \mathbb{E}_{f_i^r}\left[\log\left(1 - D(f_i^r)\right)\right].$$



Optimal-Transport Label Assignment

- An algorithm aims to transport unlabeled infrared samples to the generated visible pseudo classes. Let $\boldsymbol{P}^r \in \mathbb{R}^{N_r \times N_p}$ denote softmax output of classifier for infrared data. We use $\boldsymbol{P}^r$ to act as cost measuring the difficulty of each data assigned to the identity:

$$\min_{\boldsymbol{Q}^r} \langle \boldsymbol{Q}^r, -\log(\boldsymbol{P}^r)\rangle + \frac{1}{\delta} KL(\boldsymbol{Q}^r || \boldsymbol{\alpha}\boldsymbol{\beta}^T).$$

$$\text{s.t.} \begin{cases} \boldsymbol{Q}^r \mathbb{1} = \boldsymbol{\alpha}, \boldsymbol{\alpha} = \mathbb{1} \cdot \frac{1}{N_r}, \\ \boldsymbol{Q}^{rT} \mathbb{1} = \boldsymbol{\beta}, \boldsymbol{\beta} = \mathbb{1} \cdot \frac{1}{N_p}, \end{cases}$$

where $\boldsymbol{Q}^r \in \mathbb{R}^{N_r \times N_p}$ represents the plan used for pseudo label assignment. The optimal solution $\boldsymbol{Q}^r$ can be achieved by the iteratively Sinkhorn-Knopp algorithm:

$$\forall i: \ \boldsymbol{\alpha}_i \leftarrow \left[(\boldsymbol{P}^r)^\delta \boldsymbol{\beta}\right]_i^{-1} \quad \forall j: \ \boldsymbol{\beta}_j \leftarrow \left[\boldsymbol{\alpha}^T (\boldsymbol{P}^r)^\delta\right]_j^{-1},$$

$$\boldsymbol{Q}^r = \mathrm{diag}(\boldsymbol{\alpha})(\boldsymbol{P}^r)^\delta \mathrm{diag}(\boldsymbol{\beta}).$$

Prediction Alignment Learning

- A loss function aims to alleviate incorrect assignment. Let $\boldsymbol{S}^v \in \mathbb{R}^{B \times N_p}$, $\boldsymbol{S}^r \in \mathbb{R}^{B \times N_p}$ denote visible and infrared predictions. We firstly mix intra-/cross-modality prediction by self-attention then compute KL-divergence between them:

$$\boldsymbol{S}^{vr} = \mathrm{softmax}\left(\boldsymbol{S}^v (\boldsymbol{S}^r)^T\right)\boldsymbol{S}^r, \qquad \boldsymbol{S}^{rr} = \mathrm{softmax}\left(\boldsymbol{S}^r (\boldsymbol{S}^r)^T\right)\boldsymbol{S}^r,$$

$$\boldsymbol{S}^{rv} = \mathrm{softmax}\left(\boldsymbol{S}^r (\boldsymbol{S}^v)^T\right)\boldsymbol{S}^v.$$

$$\mathcal{L}_a^{vr} = KL(\boldsymbol{S}^v || \boldsymbol{S}^{vr}), \mathcal{L}_a^{rv} = KL(\boldsymbol{S}^{rv} || \boldsymbol{S}^{rr}), \quad \mathcal{L}_a = \lambda_a^{vr}\mathcal{L}_a^{vr} + \lambda_a^{rv}\mathcal{L}_a^{rv}.$$

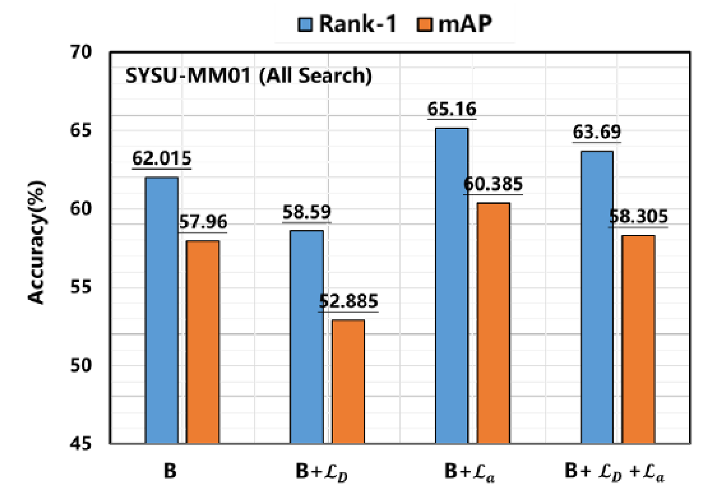
III. Experiments

Main results on SYSU-MM01 and RegDB

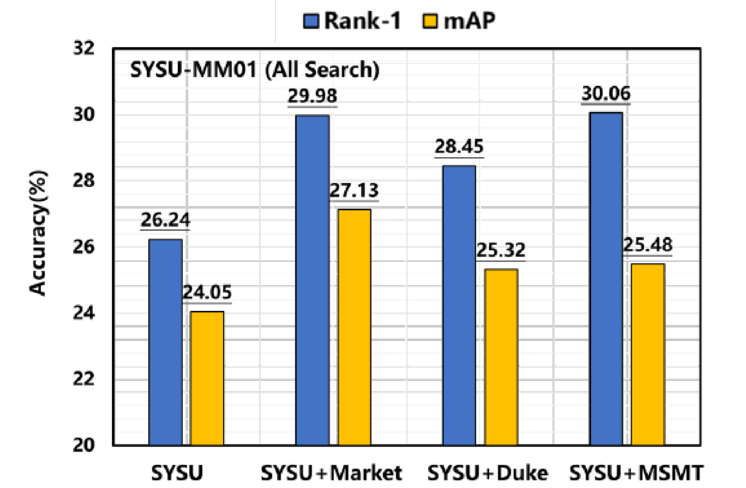
Settings			SYSU-MM01				RegDB			
			All Search		Indoor Search		Visible2Thermal		Thermal2Visible	
Type	Method	Venue	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
UDA-ReID	SSG‡[9]	ICCV'19	2.3	12.7	-	-	2.2	2.9	-	-
	ECN‡[48]	CVPR'19	8.1	5.0	-	-	1.9	3.2	-	-
	D-MMD†[21]	ECCV'20	12.5	10.4	19.0	15.4	2.2	3.7	2.0	3.6
	MMT†[11]	ICLR'20	13.9	8.4	21.0	15.3	5.3	7.1	11.0	12.1
	SpCL(UDA)†[12]	NIPS'20	15.1	6.5	19.5	12.1	3.3	4.3	8.4	9.5
	GLT†[44]	CVPR'21	7.7	9.5	12.1	18.0	2.9	4.5	6.3	7.6
USL-ReID	BUC†[17]	AAAI'19	8.2	3.2	12.5	6.0	4.7	4.5	8.8	6.0
	SpCL(USL)†[12]	NIPS'20	18.7	11.4	27.1	20.9	20.6	17.3	19.0	16.6
	MetaCam†[36]	CVPR'21	14.7	9.3	23.9	17.1	23.1	17.5	20.9	16.5
	HCD†[46]	ICCV'21	18.0	17.9	24.4	28.8	10.8	12.3	12.4	13.7
SVI-ReID	JSIA-ReID[29]	AAAI'20	38.1	36.9	43.8	52.9	48.5	49.3	48.1	48.9
	Hi-CMD[3]	CVPR'20	34.9	35.9	-	-	70.9	66.0	-	-
	AGW[39]	TPAMI'21	47.5	47.7	54.17	63.0	70.1	66.4	70.5	65.9
	NFS[2]	CVPR'21	56.9	55.5	62.8	69.8	80.5	72.1	78.0	69.8
	LbA[24]	ICCV'21	55.4	54.1	58.5	66.3	74.2	67.6	72.4	65.5
	CAJL[37]	ICCV'21	69.9	66.9	76.3	80.4	85.0	79.1	84.8	77.8
	MPANet[35]	CVPR'21	70.6	68.2	76.7	81.0	83.7	80.9	82.8	80.7
USVI-ReID	H2H[16]	TIP'21	25.5	25.2	-	-	14.1	12.3	13.9	12.7
	Ours	-	**29.9**	**27.1**	**29.8**	**38.8**	**32.9**	**29.7**	**32.1**	**28.6**
SSVI-ReID	Ours	-	**48.2**	**43.9**	**47.4**	**56.8**	**49.9**	**41.8**	**49.6**	**42.8**

Ablation Study on SYSU-MM01

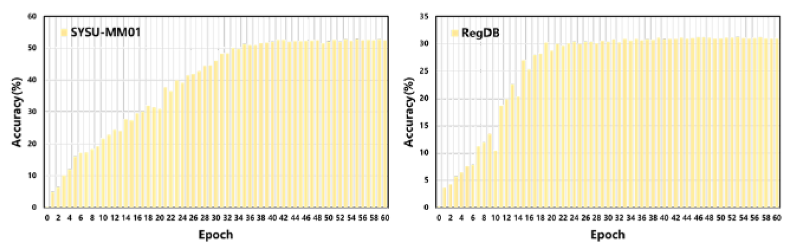
Order	Approach				All Search							
					USVI-ReID				SSVI-ReID			
	$\mathcal{L}_{V\text{-}ReID}$	$\mathcal{L}_D$	$\mathcal{L}_a$	$OTLA$	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
1	✓	-	-	-	12.62	41.91	57.27	12.73	12.25	46.49	62.24	14.66
2	✓	✓	-	-	16.62	49.91	64.53	15.94	23.69	59.56	73.10	24.71
3	✓	-	✓	-	12.65	42.39	57.03	12.81	13.86	46.67	61.87	14.72
4	✓	✓	-	✓	20.90	59.53	73.86	19.83	33.89	73.89	85.49	32.44
5	✓	-	✓	✓	19.64	61.16	77.31	19.74	36.31	77.31	86.93	34.66
6	✓	✓	✓	✓	29.98	71.79	83.85	27.13	48.15	85.30	92.64	43.86



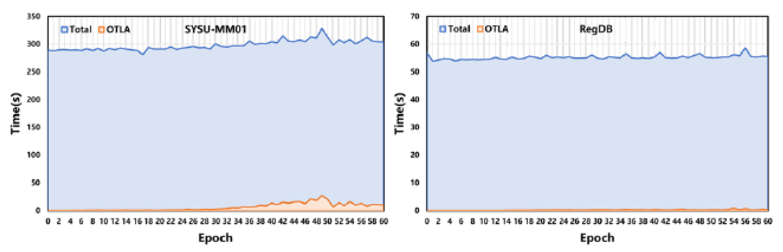
Supervised Setting for $\mathcal{L}_D$ and $\mathcal{L}_a$



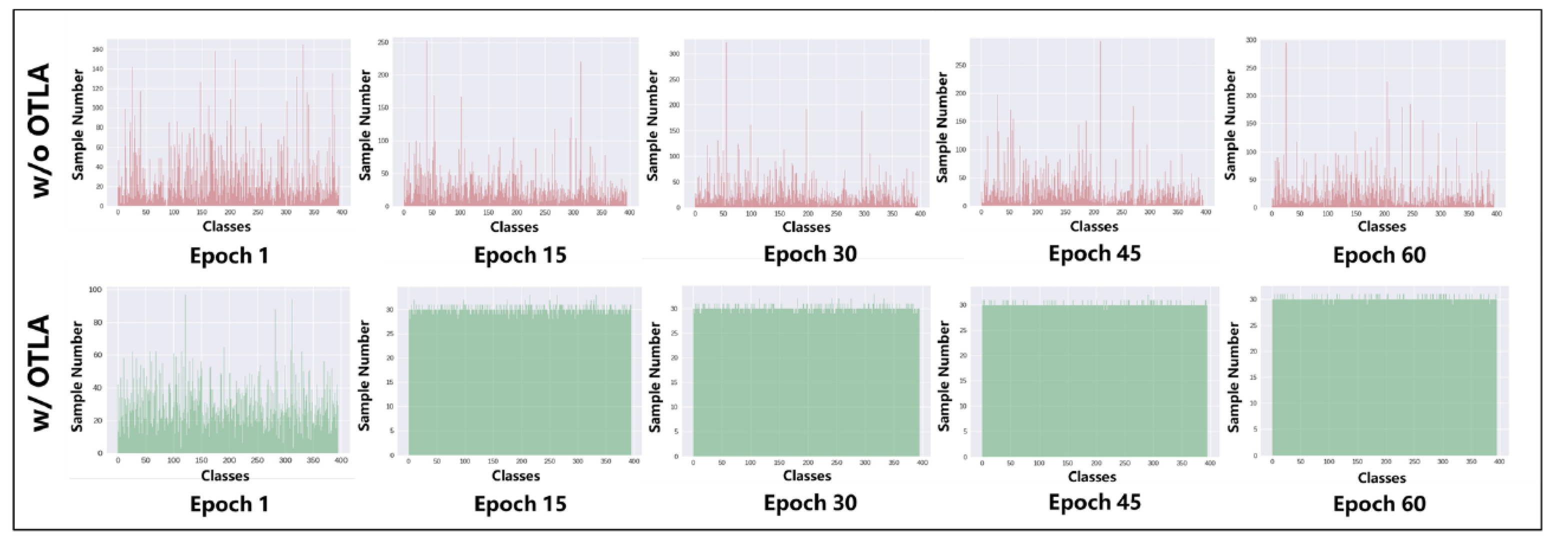
Effects of Visible Source Domain



(a) Label accuracy of SYSU-MM01 and RegDB



(b) OTLA time analysis of SYSU-MM01 and RegDB



(c) Infrared pseudo label distribution of SYSU-MM01