Disentangling based Environment-Robust Feature Learning for Person ReID

Yifan Liu, Ya-Li Li, Shengjin Wang

Department of Electronic Engineering Tsinghua University ,Beijing,China

Abstract

Person re-identification suffers from the influence of environments, as a cross-camera retrieving problem. Images captured by the same camera have similar backgrounds, illuminations and angles, making them seem similar, even if their person IDs are different. A quantitative experiment is designed in this paper to demonstrate the above issue. We proposed a novel Environment-Robust Feature Learning network (EFL) to tackle this problem. First, we designed a feature disentangling module (FDM) based on the idea of minimizing mutual information of identity related features and camera related features. Besides, we adopt a Mutual Mean Teaching (MMT) framework as identity feature extractor to improve the robustness of the features. Moreover, we constructed a multi-environment person ReID dataset ME-ReID (multi-environment) to evaluate our method, which contains more complicated environment variations comparing to existing datasets.

Introduction

Person re-identification aims at retrieving a person across different cameras. Images captured by the same camera have similar environmental features such as background, illumination, and angle, Resulting in that images of different people captured by the same camera may have high similarity scores, which would distract the feature matching and degrade the ReID performance.

Table. 1 shows the quantitative evaluations of this issue. We calculate the proportion of each kind of valid top-1 retrieval results on several person ReID benchmarks. Results of baseline method show that the proportion of wrong retrievals under same camera exceeds 50% on all of the benchmarks. Results also show that, our purposed method eliminates environment related factors from extracted features, reducing the similarities between images with different identities but same camera source, correcting an amount of wrong retrievals under the same camera.

dataset	method	S. ID D. Cam	D. ID S. Cam	D. ID D. cam	prop.
Market1501[31]	baseline	94.4	2.9	2.7	51.8
	ours	**95.0**	**2.4***(-0.5)*	**2.6**	**48.0***(-3.8)*
MARS[32]	baseline	89.9	6.8	3.3	67.3
	ours	**91.2**	**5.5***(-1.3)*	**3.3**	**62.5***(-4.8)*
MSMT17[25]	baseline	80.0	13.2	6.8	66.0
	ours	**81.0**	**11.8***(-1.4)*	**7.2**	**62.1***(-3.9)*
ME-ReID	baseline	60.9	21.5	17.6	56.0
	ours	**66.9**	**15.4***(-6.1)*	**17.7**	**46.5***(-9.5)*

Table 1: top1 retrieval results of baseline and our method. ’S.’ and ’D.’ refers to same and different respectively. The green label represents right retrieval results, while red labels represent the wrongs. ’Prop.’ refers to the proportion of wrong retrievals under same camera in all wrong retrievals.

Materials & Methods

Methods

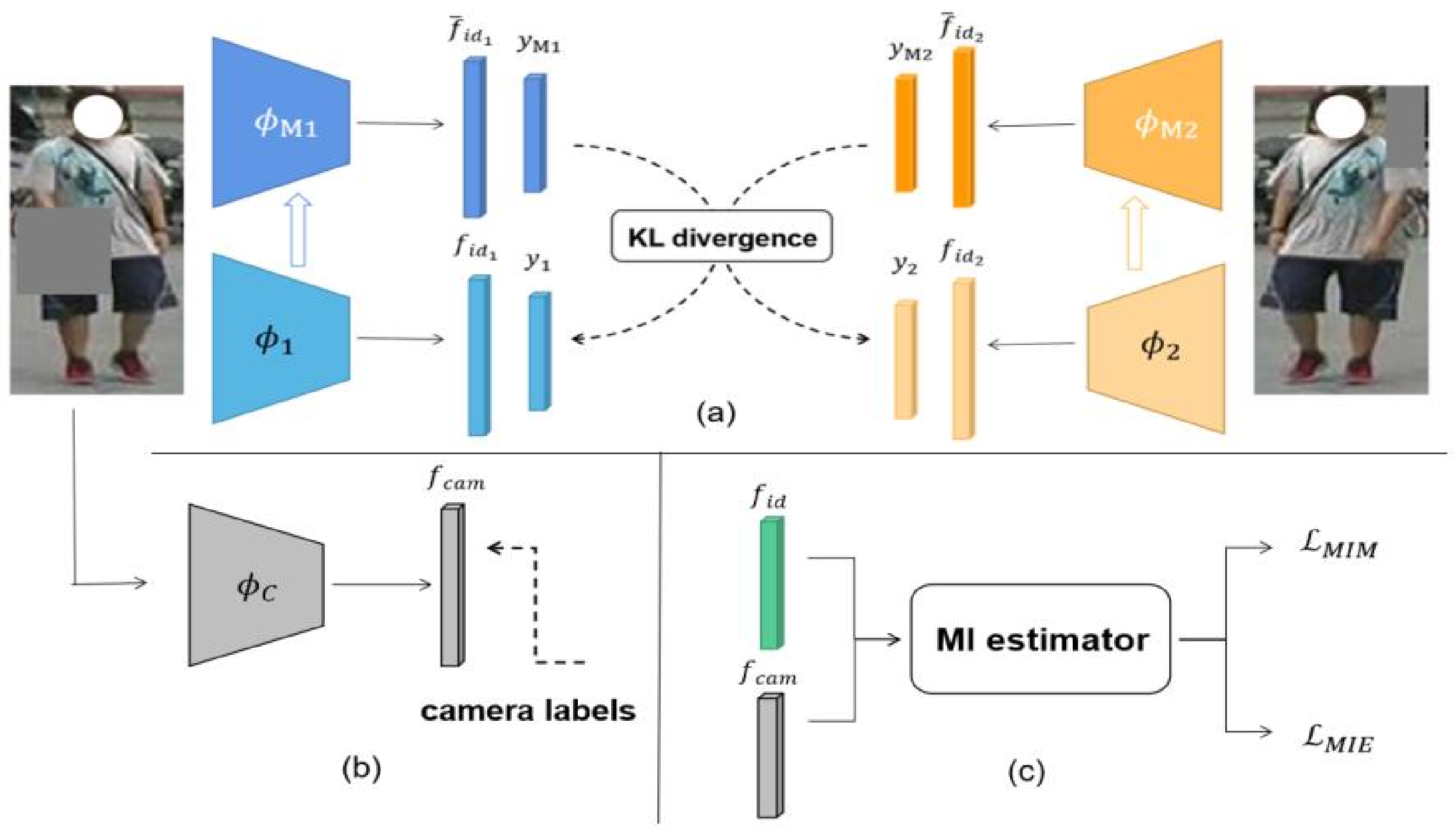

Figure 1: An overview of EFL. (a) Identity stream: Mutual Mean Teaching based identity feature learning. (b) Environment stream: using camera labels to learn discriminative environment features. (c) Feature Disentangling Module: minimizing the mutual information of identity related and camera related features.

Materials & Methods

Our proposed Environment-robust Features Learning (EFL) is a dual stream learning framework, as shown in fig. 1.

- Identity Stream: a Mutual Learning based framework is adopted to extract environment-robust features. We adopt random augmentation twice to mimic environment variations, and add KL divergence as constraint to improve the robustness to environment changes.

$$\mathcal{L}_{\mathrm{id}} = (1-\lambda_{\mathrm{KL}})\mathcal{L}_{\mathrm{ce}} + \mathcal{L}_{\mathrm{tri}} + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}$$

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{KL}(y_{\mathrm{M1}}||y_2) + \mathrm{KL}(y_{\mathrm{M2}}||y_1)$$

- Environment Stream: we train an encoder with the given camera labels, to learn accurate and discriminative environment features.
- Feature Disentangling Module: we propose a mutual information minimization module to further disentangle the environment related factors from the extracted features. Mutual information between identity features and environment features is minimized.

$$\mathrm{I(X;Y)} = \int p(x,y)log\frac{p(x,y)}{p(x)p(y)}dxdy$$

$$\begin{aligned}\hat{\mathrm{I}}(f_{\mathrm{id}};f_{\mathrm{cam}}) :=& \mathbb{E}_{p(f_{\mathrm{id}},f_{\mathrm{cam}})}[\log q(f_{\mathrm{cam}}|f_{\mathrm{id}})] \\ &- \mathbb{E}_{p(f_{\mathrm{id}})}\mathbb{E}_{p(f_{\mathrm{cam}})}[\log q(f_{\mathrm{cam}}|f_{\mathrm{id}})]\end{aligned}$$

$$\mathcal{L}_{\mathrm{MIM}} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}[\log q(f^i_{\mathrm{cam}}|f^i_{\mathrm{id}}) - \log q(f^j_{\mathrm{cam}}|f^i_{\mathrm{id}})]$$

ME-ReID Dataset

ME-ReID dataset is a new ReID dataset which contains more complicated environment variations comparing to existing datasets.

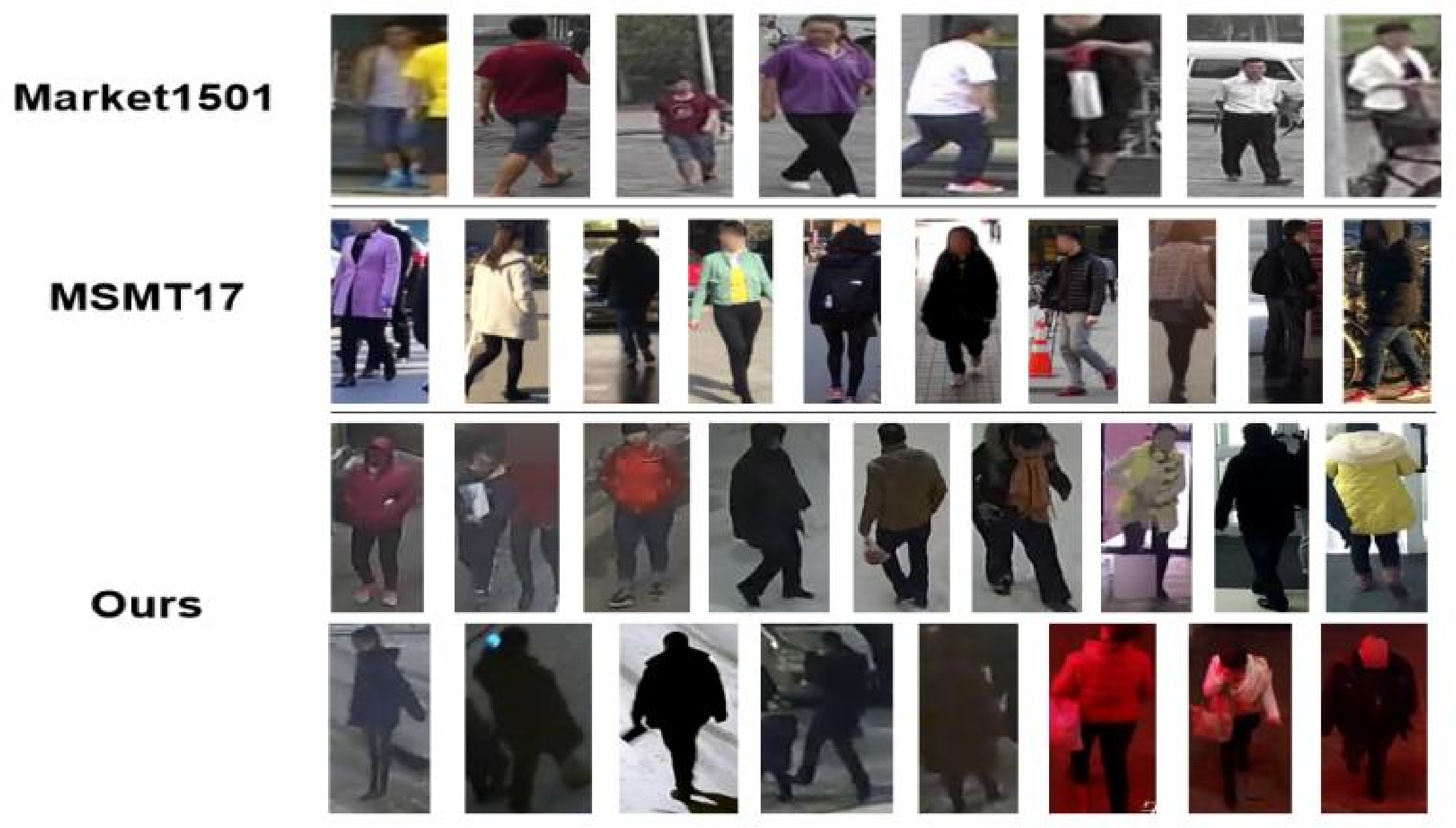

Figure 2: Examples of existing datasets and our dataset. The first row of ’Ours’ shows daytime samples, with sunny samples on the left, snowy in the middle and indoors ones on the right. The second row shows night samples, with the three on the right contain unnatural light impact.

Results

methods	Market1501		MSMT17		ME-ReID		methods	MARS	
	R1	mAP	R1	mAP	R1	mAP		R1	mAP
ResNet50	94.4	86.0	80.0	56.2	60.9	48.5	ResNet50	89.5	85.6
ResNet50+ours	**95.0**	**87.6**	**81.0**	**58.4**	**66.9**	**55.7**	**ResNet50+ours**	**90.5**	**87.0**
ResNeSt50	95.7	89.7	85.6	66.9	64.8	55.9	AP3D-NL	89.9	86.8
ResNeSt50+ours	**96.0**	**91.0**	**86.3**	**68.7**	**73.9**	**62.8**	**AP3D-NL+ours**	**91.2**	**88.0**

Table 3: Compare with baseline method of different backbones.

source	target	method	R1	mAP
MSMT17	Market1501	ResNet50	54.5	28.5
		+ours	**56.0**	**30.5**
MSMT17	ME-ReID	ResNet50	43.1	31.2
		+ours	**49.6**	**35.7**
Market1501	ME-ReID	ResNet50	33.0	22.6
		+ours	**34.7**	**24.3**

Table 5: Cross domain evaluations. The models are trained on source domain and directlytested on target domain.

Results

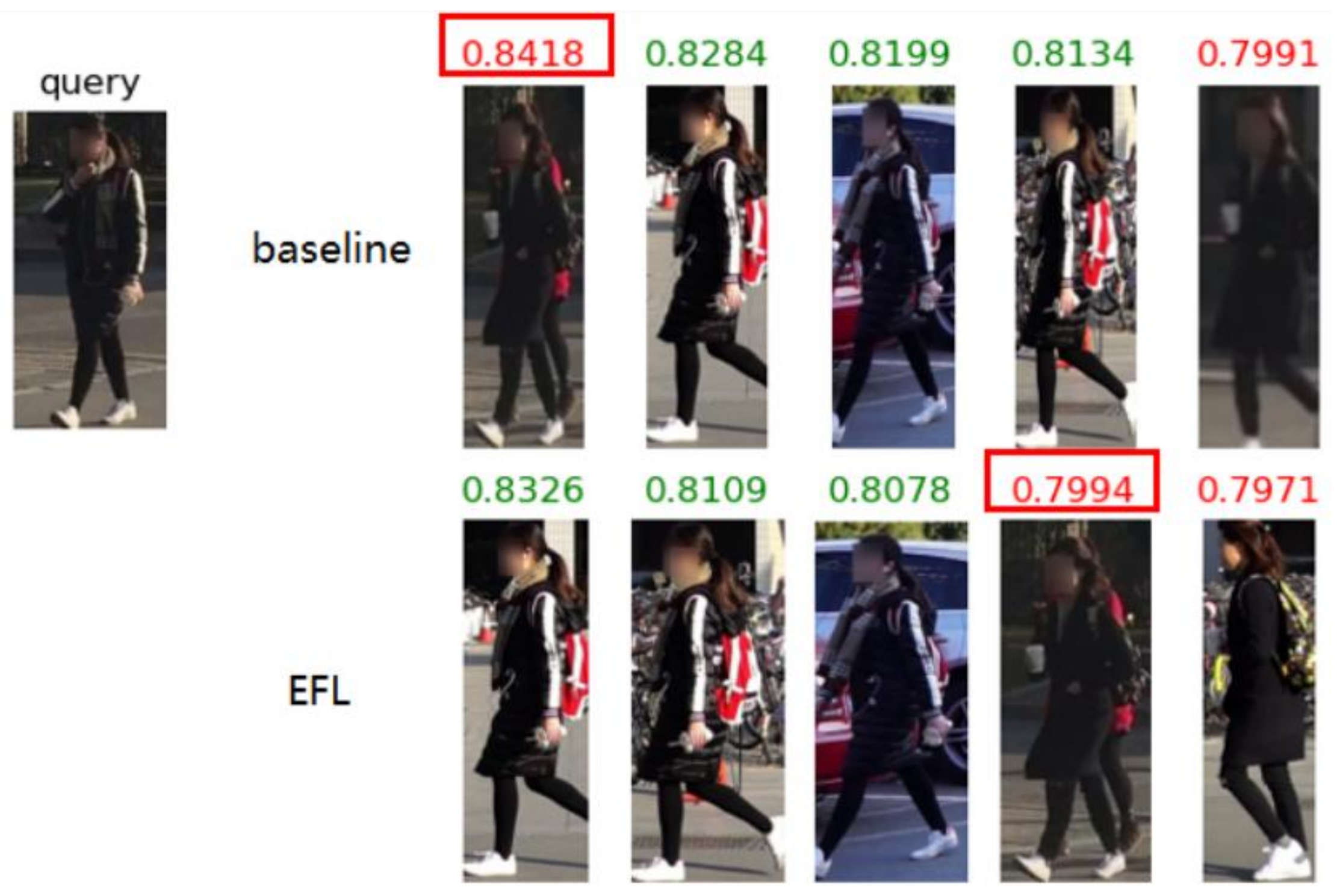

Figure 3: An retrieving example. The first row is the results of baseline, the second is the results of EFL. The number on the top of each image refers to the cosine similarity with query.

Figure 3 demonstrates that our proposed EFL method disentangles environment factors from extracted features and corrects some wrong retrieval results under same camera.

method	MMT	FDM	MSMT17		MARS	
			R1	mAP	R1	mAP
baseline	×	×	80.0	56.2	89.5	85.6
	×	✓	80.7	56.6	90.3	86.0
	✓	×	80.4	58.0	90.2	86.4
EFL	✓	✓	**81.0**	**58.4**	**90.5**	**87.0**

Table 4: Ablation studies on MSMT17 and MARS datasets, with the ResNet50 backbone.

Methods	Backbone	Market1501		MSMT17	
		R1	mAP	R1	mAP
*CBN [37]	ResNet50	91.3	77.3	72.8	42.9
PCB+RPP [19]	ResNet50	93.8	81.6	-	-
MGN [22]	ResNet50	**95.7**	86.9	76.9	52.1
HOReID [21]	ResNet50	94.2	84.9	-	-
DGNet [33]	ResNet50	94.8	86.0	77.2	52.3
***EFL(ours)**	ResNet50	95.0	**87.6**	**81.0**	**58.4**
RGA-SC [30]	R50-RGA	**96.1**	88.4	80.3	57.5
ABD-Net [3]	ABD-Net	95.6	88.3	82.3	60.8
BAT-net [7]	BAT-net	95.1	87.4	79.5	56.8
OSNet [35]	OSNet	94.8	84.9	78.7	52.5
OP-ReID [26]	R50-DNL	**96.1**	89.3	-	-
*TransReID [12]	ViT-B/16	95.2	88.9	85.3	67.4
***EFL(ours)**	ResNeSt50	**96.0**	**91.0**	**86.3**	**68.7**

Table 5: Compare with state-of-the-art methods on large-scale image-level ReID benchmarks Market1501 and MSMT17. ‘*’ denotes using camera labels.

Methods	MARS		
	R1	R5	mAP
GLTR [16]	87.0	95.8	78.5
STE-NVAN [17]	88.9	-	81.2
AFA [2]	90.2	96.6	82.9
GRL [18]	91.0	96.7	84.8
STT [28]	88.7	-	86.3
DenseIL [13]	90.8	97.1	87.0
PSTA [24]	**91.5**	-	85.8
SINet [1]	91.0	-	86.2
EFL(ours)	91.2	**97.8**	**88.0**

Table 6: Comparing with video ReID methods on MARS.

Conclusion

We proposed a novel Environment-robust Feature Learning (EFL) network to eliminate the interfere of environment related features among each camera. EFL is a multi-task learning framework, extracting disentangled identity related features and environment related features by minimizing the mutual information between them. Besides, we proposed a new ReID dataset ME-ReID with complicated environment varieties, which is more challenging and closer to practical application scenarios. Extensive experiments demonstrated the effectiveness of our proposed EFL method.

Contact Information
liu-yf21@mails.Tsinghua.edu.cn