Disentangling based Environment-Robust Feature Learning for Person ReID

Yifan Liu
liu-yf21@mails.tsinghua.edu.cn
Yali Li
liyali13@tsinghua.edu.cn
Shengjin Wang
wgsgj@tsinghua.edu.cn

Department of Electronic Engineering,
Tsinghua University, China

Abstract

Person re-identification (ReID) has received much attention in recent years. As a cross-camera retrieving problem, ReID suffers from the influence of environments. Images captured by the same camera have similar environment features like backgrounds, illuminations and angles, making their features extracted by a neural network have high similarity scores, 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. Extensive experiments show that our method achieves state-of-the-art performances on widely used datasets Market1501, MSMT17, MARS. Our method also has a great improvement of +9.1%/+6.9% of rank1/mAP on ME-ReID, showing the effectiveness of our method. The ME-ReID dataset is available on: https://github.com/liuyf21/ME-ReID-dataset.

1 Introduction

Person re-identification (ReID), which aims at pedestrian retrieval in a cross-camera setting, is an important issue in computer vision. The key issue of person ReID is to extract discriminative features. That is, the features between the same person should have higher similarity scores than features between different people. However, images captured by the same camera have similar environmental features such as background, illumination, and angle, which are independent of pedestrian identities. These identity-irrelevant environmental features result in that images 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. In contrast, negative samples from the same camera only account for 1/N of all negative samples on average, which is far less than 50%. N is the number

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.

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)*

of cameras in the gallery set, equaling to 6, 6, 15 and 30 on Market1501[31], MARS[32], MSMT17[25] and our proposed ME-ReID respectively. This shows that although the model is trained to fit the identity labels, camera related environmental features still interfere with the process of ReID. The results in table. 1 also show that, our purposed method eliminates environment related factors from the identity features, reducing the similarities between images with different identities but same camera source, correcting an amount of wrong retrievals under the same camera.

Existing methods cannot solve the environment problem well. Some works use human parsing to focus on foreground parts[20, 36], but only eliminate background variation without considering illuminations and angles. Other works utilize camera labels to solve environment issues[10, 12, 37], but cannot eliminate environment features well. Camera-based Batch Normalization (CBN)[37] is proposed to force the images from different cameras to have similar distributions. He et. al. inserted Side Information Embeddings(SIE)[12] into transformer encoders and Hao et. al. proposed a camera-aware center loss[10]. Different from the existing works, our proposed EFL learns environment-robust features in a feature disentangling way, based on the idea of mutual information minimizing.

In this work, we propose a novel Environment-robust Features Learning Network (EFL) to extract identity discriminative pedestrian features. EFL is a dual stream multi-task learning framework consisting an identity stream and an environment stream. For identity stream, we adopt a Mutual Mean Teaching [7] based feature extractor to learn robust identity features. For environment stream, we train a camera encoder using camera labels to learn accurate environment features. In particular, to eliminate the camera related parts in identity features, we design a Feature Disentangling Module (FDM) based on mutual information minimization.

We conduct extensive experiments on existing large-scale benchmarks, including image-level ReID datasets Market1501, MSMT17 and video-level dataset MARS. Furthermore, to thoroughly investigate the environment issues in person ReID, we construct a new ReID dataset ME-ReID (multi-environment) for evaluation and benchmarking. Existing datasets are with simple and idealistic environments, making them hard to adapt to complex real-world scenes. In contrast, our new dataset contains abundant environment varieties, including day and night, weather and illumination changes, indoor and outdoor, etc. Because of these environment varieties, ME-ReID is more challenging and more suitable to evaluate how a method extracts environment-robust features.

In general, our work consists of three folds: 1) We propose a novel Environment-robust

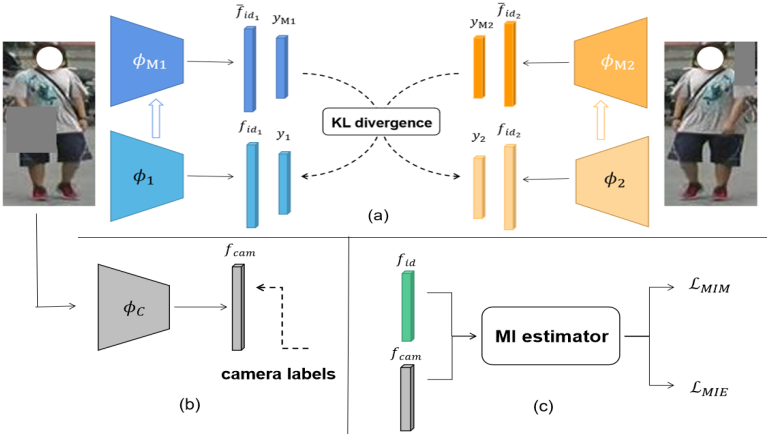

Figure 1: An overview of EFL. (a) Identity stream: Mutual Mean Teaching based identity feature learning. (b) Environment stream: using camera labels to learn discriminative environment features. (c) Feature Disentangling Module: minimizing the mutual information of identity related and camera related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

Features Learning Network, which disentangles environment related factors from identity features by mutual information minimization. 2) We constructed a multi-environment ReID dataset ME-ReID. Comparing with existing datasets, our dataset has more complicated environments, thus closer to particular applications. 3) Extensive experiments demonstrate the effectiveness of our method on several widely used datasets Market1501[31], MSMT17[25] and MARS[32]. ME-ReID is also used to evaluate our method.

2 Method

Our proposed Environment-robust Features Learning (EFL) is a dual stream learning framework, as shown in fig. 1. For identity stream, we design a feature extraction module based on Mutual Mean Teaching (MMT) [8] to learn robust identity related features. For environment stream, we train an encoder with the given camera labels, to learn accurate and discriminative environment features. After the dual stream feature extracting, we adopt a mutual information minimization based Feature Disentangling Module (FDM) to further eliminate the interfere of camera environment related parts in the identity features.

2.1 Identity related Feature Learning

As shown in Fig. 1(a), we adopt a MMT based framework to learn features which are discriminative and robust for environment changes. Mutual learning[29] can be used to extract invariant features of the two networks with the constraint of KL divergence. Inspired by this, we mimic two different environment by randomly augment an image twice and separately input them into a network. In this way, the extracted features are more robust to environment changes. Besides, we adopt mean nets to avoid the two networks converging to equal to each other and loss their independence, following [8].

We denote the input image as x and the augmented ones as x_1 and x_2. Augmented images

$x_{1,2}$ are input to two neural nets $\phi_{1,2}$ with parameters denoted as $\theta_{1,2}$. The data augmentations are used to mimic environment changes, including random horizontal flip for angle variations and random erasing for background changes and occlusions.

We use Cross Entropy Loss as classification loss and Triplet Loss as metric learning loss. We also adopt KL divergence for invariant feature learning between the two networks. The loss function to optimize ϕ_1 and ϕ_2 is denoted as:

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

where $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{tri}}$ are separately calculated for ϕ_1 and ϕ_2. λ_{KL} is a hyperparameter. To avoid collapse between ϕ_1 and ϕ_2, we adopt additional mean nets $\phi_{\text{M1,M2}}$ with parameters $\mathbb{M}[\theta_{1,2}]$ as teacher models. $\phi_{\text{M1,M2}}$ are optimized by eq. 2.

$$\mathbb{M}^T[\theta_{1,2}] = \alpha\mathbb{M}^{T-1}[\theta_{1,2}] + (1-\alpha)\theta_{1,2} \tag{2}$$

where the superscripts denote the training iterations, α is a momentum hyperparameter.

$\mathcal{L}_{\text{KL}}$ is the KL divergence between the predictions of teacher model and student model, which is formulated as eq. 3.

$$\mathcal{L}_{\text{KL}} = \text{KL}(y_{\text{M1}}||y_2) + \text{KL}(y_{\text{M2}}||y_1) \tag{3}$$

By adding this constraint, predictions of differently augmented inputs are encouraged to be equal. In this way, the identity encoders are trained to be robust to different environment variations. Thus, the extracted features suffer less from environment noise and have better performance.

2.2 Feature Disentagling Module

In order to remove the interfere of camera related factors, like background and illumination, we design a feature disentangling module to learn nearly independent identity features and camera environment features, denoting as f_{id} (extracted from ϕ_1 or ϕ_2) and f_{cam}. We take advantage of the camera labels to train a camera encoder ϕ_c to extract accurate f_{cam}, as shown in fig. 1(b). Cross Entropy loss and Triplet loss are adopted to optimize ϕ_c.

The mutual information of f_{id} and f_{cam} is calculated and minimized. By doing so, the f_{id} is encouraged to be invariant to different camera environments, while preserving identity related factors.

Mutual information is a statistic to measure the dependence between two random variables. The MI between random variables X and Y is formulated as:

$$\text{I(X;Y)} = \int p(x,y)log\frac{p(x,y)}{p(x)p(y)}dxdy \tag{4}$$

where $p(x,y)$ denotes the joint probability distribution, $p(x)$ and $p(y)$ denote the marginal ones. Mutual information strictly quantifies the information between random variables, even if the variables have nonlinear dependence. So it measures true mutual dependence between variables comparing with correlations [15].

As it is hard to obtain formulaic representations of the distributions of f_{id} and f_{cam}, a mutual information estimation approach is needed. We follow the implementation of Contrastive Log-ratio Upper Bound (CLUB) [4], which estimates MI as the difference of conditional probability between positive and negative pairs:

$$\begin{aligned}\hat{\text{I}}(f_{\text{id}};f_{\text{cam}}) :=& \mathbb{E}_{p(f_{\text{id}},f_{\text{cam}})}[\log q(f_{\text{cam}}|f_{\text{id}})] \\ &- \mathbb{E}_{p(f_{\text{id}})}\mathbb{E}_{p(f_{\text{cam}})}[\log q(f_{\text{cam}}|f_{\text{id}})]\end{aligned} \tag{5}$$

where $q(f_{\text{cam}}|f_{\text{id}})$ is an approximated conditional distribution modeled by a MI estimator Q. In details, the network Q estimates the mean and variance of the conditional distribution $p(f_{\text{cam}}|f_{\text{id}})$, and then derives to the approximated distribution $q(f_{\text{cam}}|f_{\text{id}})$ with the assumption that $p(f_{\text{cam}}|f_{\text{id}})$ follows a Gaussian distribution.

Derived from eq. 5, the mutual information minimization loss is formulated as:

$$\mathcal{L}_{\text{MIM}} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}[\log q(f_{\text{cam}}^{i}|f_{\text{id}}^{i}) - \log q(f_{\text{cam}}^{j}|f_{\text{id}}^{i})] \tag{6}$$

where N denotes the training batch size. We optimize identity encoders using eq. 6 as MI minimization constraint. By doing so, the extracted identity features are encouraged to be independent to environment features, which improves the performance.

For accurate MI estimating, Q is optimized by minimizing the KL divergence between the real and approximated conditional distribution:
$\text{KL}(p(f_{\text{cam}}|f_{\text{id}})||q(f_{\text{cam}}|f_{\text{id}}))$, which can be derived to the following equation.

$$\mathcal{L}_{\text{MIE}} = -\mathbb{E}_{p(f_{\text{id}},f_{\text{cam}})}[\log q(f_{\text{cam}}|f_{\text{id}})] \tag{7}$$

2.3 Joint Optimization

As shown in fig. 1, there are totally six models need to be trained in our proposed EFL framework, including identity encoders: ϕ_1, ϕ_2, ϕ_{M1}, ϕ_{M2}, a camera encoder ϕ_c and a MI estimator Q. We disentangle camera environment related part from identity feature by minimizing mutual information between f_{id} and f_{cam}. As ϕ_{M1} and ϕ_{M2} are not trained in a gradient descending way, $\mathcal{L}_{\text{MIM}}$ in eq. 6 is only used for optimization of ϕ_1 and ϕ_2. ϕ_c is trained with camera labels, adopting Cross Entropy Loss and Triplet Loss as loss functions. For stability, we train ID Nets and MI estimator alternatively in each training iteration. The algorithm of the training process is summarized as algorithm 1.

Algorithm 1: the training process of EFL

Input: batch of training images x and augmented ones x_1, x_2
hyperparameters λ_{KL}, λ_{MI}, α, N_{MI};

initialize parameters of ID Nets, ID Mean Nets, Camera Net and MI estimator.;
for *each training iter* **do**
  Identitiy encoders foward;
  calculate ID loss by eq. 1;
  Camera encoder forward: $f_{\text{cam}} = \phi_c(x_1)$;
  MI estimator forward and calculate $\mathcal{L}_{\text{MIM}}$ by eq. 6;
  combinine eq. 1 and eq. 6: $\mathcal{L}_{\text{id}} + \lambda_{\text{MI}}\mathcal{L}_{\text{MIM}}$ and optimize ϕ_1 and ϕ_2.;
  optimize ID mean nets ϕ_{M1} and ϕ_{M2} by eq. 2.;
  optimize Camera Net.;
  for *i=1 to* N_{MI} **do**
    calculate $\mathcal{L}_{\text{MIE}}$ and optimize MI estimator by eq. 7;
  end
end

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.

3 Experiments

3.1 Datasets & Evaluation Protocols

We conduct the experiments on several widely-used large-scale ReID datasets, including image-level Market1501 [31] and MSMT17 [25], and video-level MARS [32] to evaluate the performance of EFL. The evaluation protocols are mAP and CMC following [31].

We also evaluate our method on our proposed ME-ReID. As shown in fig. 2, our dataset contains more environment variations comparing to Market1501 and MSMT17, including different time (day or night), different weather(sunny or snowy), indoor or outdoor and unnatural light. The statistics of the ReID datasets are shown in table 2. ME-ReID totally contains 5908 person boundingboxes of 370 pedestrians, collected from 30 cameras (27 outdoors and 3 indoors). Though ME-ReID is smaller in scale comparing to existing large-scale datasets, it still has research value because of its environment diversities that existing datasets do not contain.

Table 2: Statistics comparing with existing datasets. I or V denotes image-level or video-level, Out. Cam. and In. Cam. denote the number of outdoor cameras and indoor cameras.

Dataset	Market[31]	MSMT[25]	CAVIAR[6]	MARS[32]	iLID-VID[23]	**ME-ReID**
I or V	I	I	I	V	V	**I**
BBoxes	32668	126411	610	1067516	42460	**5908**
Clips	-	-	-	20715	600	**-**
Identities	1501	4101	72	1261	300	**370**
Out. Cam.	6	12	0	6	0	**27**
In. Cam.	0	3	2	0	2	**3**

3.2 Settings

All the input images are resized into 256×128. During training, random horizontal flipping and random erasing[34] are adopted as augmentations. In video-level ReID settings, each frame of the input clip is ensured to be equally augmented in order not to spoil inter-frame

information. We adopt ResNet50 [11] and ResNeSt50 [27] as backbones for image-level ReID, ResNet50 and AP3D-NL[9] for video-level ReID. For the training of ME-ReID, as the training set is not large enough, we use models pretrained on large-scale ReID dataset MSMT17 for initialization, to obtain stable performance. For the training of other datasets, the backbones are initialized with models pretrained on ImageNet [5].

For image-level ReID, each mini-batch consists of 16 identities, with 16 instances of each identity on MSMT17, 8 instances on Market1501, and 4 instances on ME-ReID. For video-level ReID, each mini-batch consists of 12 identities, with 4 clips of each identity, each video clip contains 4 frames, and temporal average pooling is used to fuse the frames. We train totally 240 epochs for all the datasets. The optimizer is ADAM [14] with weight decay 0.0005 adopted to regularize the parameters. The initial learning rate of ID Nets and Camera Net is 0.0003, with multiplied by 0.1 at the 60, 120, 180 epochs. The initial learning rate of MI estimator is 0.00001, equally decreased with other networks. As for the hyperparameters, α, λ_{MI}, λ_{KL}, N_{MI} is set to 0.999, 0.0001, 0.5, 10 respectively.

3.3 Compare with Baseline Models

We test ResNet50 and ResNeSt50 backbones on image-level ReID datasets, ResNet50 and AP3D-NL on video-level dataset MARS. The results are shown in table 3. Our method obtains consistent improvements on all of the baseline models and datasets. It is worth mentioning that our method shares totally the same structures with baseline models, without any additional parameters and computations in the testing phase. Under this circumstance, the improvements are considerable, especially for the mAP scores. Our proposed EFL obtains +1.6% mAP improvement on Market1501, +1.0%/+2.2% Rank1/mAP improvement on MSMT17 and +1.0%/+1.4% Rank1/mAP on MARS with ResNet50 backbone. Besides, EFL obtains +1.3% and +1.8% mAP improvement on Market1501 and MSMT17 with ResNeSt50 backbone, and +1.3%/+1.2% mAP/Rank1 on MARS with AP3D-NL backbone, which are significant improvements on those baseline models with the performance already comparable with SOTA.

Table 3: Compare with baseline method of different backbones.

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**

Comparing the retrieving results on ME-ReID with those on existing large-scale datasets, ME-ReID is far more challenging than existing dataset, because of the complicated scenarios. Besides, on ME-ReID, our method acquires a significant improvement of +6.0%/+7.2% Rank1/mAP with ResNet50 backbone and +9.1%/+6.9% Rank1/mAP with ResNeSt50 backbone, which shows that EFL also works well on real surveillance applications. EFL achieves greater improvement on ME-ReID than other datasets, because ME-ReID has more complex scenarios, making the environment related noise interfere more severe.

Fig 3 shows examples of retrieving results. In both examples, the top-1 retrieval result of baseline is an image with different identity but same camera label, which reappear in the result of our method with lower cosine similarities (marked with red boxes in fig 3). Indicating that our method eliminates the environment related factors from identity features, making more positive samples rank ahead of negative samples under the same camera.

Figure 3: Two examples of retrieving divided by a vertical line. For each sample, the first row is the results of baseline method, the second row is the results of EFL method. The number on the top of each image refers to the cosine similarity with query.

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

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 5: Cross domain evaluations. The models are trained on source domain and directly tested on target domain.

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**

3.4 Ablation Study

We further present ablation study on MSMT17 and MARS datasets, to investigate the effects of different modules. The results of table 4 show that, separately adding MMT framework and feature disentangling module to baseline obtains an improvement. By adding MMT framework, the network learns to extract features that are invariant to identity unrelated factors, like backgrounds, occlusions and angles with the implementation of KL divergence between outputs encoded from differently augmented images, leading to improvements of 0.4% R1/1.8% mAP on MSMT17 and 0.7% R1/0.8% mAP on MARS. By adding FDM, the identity features are decoupled from environment features by minimizing the mutual information between them, which further eliminate camera environment related factors in identity features, leading to improvements of 0.7% R1 / 0.4% mAP on MSMT17 and 0.8% R1 / 0.4% mAP on MARS. Further improvement is achieved when combining Mutual Mean Teaching framework and Feature Disentangling Module.

3.5 Cross Domain Evaluations

We demonstrate the result of cross domain evaluations in table 5. The models are trained on the source dataset and directly tested on the target dataset. We evaluate our method on several cross domain settings, and it turns out that EFL obtain consistent and considerable improvement on all of the cross domain settings. On the MSMT17 to ME-ReID setting, our method achieved a significant improvement of 6.5% / 4.5% Rank1 / mAP. On MSMT17 to Market1501 and Market1501 to ME-ReID settings, EFL achieved improvements of 1.5% /

Table 6: Compare with state-of-the-art methods on large-scale image-level ReID benchmarks Market1501 and MSMT17. '*' denotes using camera labels.

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 [23]	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 7: Comparing with video ReID methods on MARS.

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**

2.0% Rank1 / mAP and 1.7% / 1.7% Rank1 / mAP, which are also considerable. As different cameras can be regarded as different domains, our method eliminate the environment differences between each camera domain and extract inherent identity discriminative features, which also leads to an improvement on cross domain ReID problems.

3.6 Compare with State-of-the-Art Methods

We compare the proposed method with the state-of-the-art on public datasets such as on Market1501, MSMT17 and MARS. The results are presented in table 6 and table 7. Our method acquires competitive performances comparing with SOTA. In table 6, we first compare with existing methods on the same backbone as ResNet50. Comparing with image-level methods using ResNet50 as backbone, our method exceeds existing method of 1.6% mAP on Market1501, and 4.8%/6.1% Rank1/mAP on MSMT17. We further adopt ResNeSt50 as backbone for fair comparisons with methods which use backbones with larger parameters. Comparing with SOTA, our proposed EFL method achieves mAP scores 1.7% higher on Market1501, 1.3% higher on MSMT17, 1.0% on MARS, which are significant improvements. Besides, our method achieves a 1.0% higher R1 score on MSMT17 and a 0.7% R5 score on MARS. Our method also achieves comparable R1 performances with SOTA on Market1501 and MARS.

4 Conclusions

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. More experiment results and discussions are shown in the supplementary material.

Acknowledgement. This study is also supported by Tsinghua University Toyota Joint Research Center for AI Technology of Automated Vehicle(TTAD 2022-07).

References

[1] Shutao Bai, Bingpeng Ma, Hong Chang, Rui Huang, and Xilin Chen. Salient-to-broad transition for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7339–7348, 2022.

[2] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In *European Conference on Computer Vision*, pages 660–676. Springer, 2020.

[3] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8351–8361, 2019.

[4] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR, 2020.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] S. C. Dong, M. Cristani, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *British Machine Vision Conference*, 2011.

[7] Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Bilinear attention networks for person retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8030–8039, 2019.

[8] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020.

[9] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *European Conference on Computer Vision*, pages 228–243. Springer, 2020.

[10] Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-modality person re-identification via modality confusion and center aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16403–16412, 2021.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15013–15022, 2021.

[13] Tianyu He, Xin Jin, Xu Shen, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Dense interaction learning for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1490–1501, 2021.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Justin B Kinney and Gurinder S Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111 (9):3354–3359, 2014.

[16] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3958–3967, 2019.

[17] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. *arXiv preprint arXiv:1908.01683*, 2019.

[18] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, and Xiaoyun Yang. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13334–13343, 2021.

[19] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018.

[20] Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. Eliminating background-bias for robust person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5794–5803, 2018.

[21] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6449–6458, 2020.

[22] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018.

[23] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. *Springer, Cham*, 2014.

[24] Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, and Dong Wang. Pyramid spatial-temporal aggregation for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12026–12035, 2021.

[25] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018.

[26] Cheng Yan, Guansong Pang, Jile Jiao, Xiao Bai, Xuetao Feng, and Chunhua Shen. Occluded person re-identification with single-scale global representations. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11855–11864, 2021. doi: 10.1109/ICCV48922.2021.01166.

[27] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

[28] Tianyu Zhang, Longhui Wei, Lingxi Xie, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. Spatiotemporal transformer for video-based person re-identification. *arXiv preprint arXiv:2103.16469*, 2021.

[29] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. Deep mutual learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[30] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 3186–3195, 2020.

[31] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[32] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European conference on computer vision*, pages 868–884. Springer, 2016.

[33] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2138–2147, 2019.

[34] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

[35] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3702–3712, 2019.

[36] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *European Conference on Computer Vision*, pages 346–363. Springer, 2020.

[37] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *European Conference on Computer Vision*, pages 140–157. Springer, 2020.