BCaR: Beginner Classifier as Regularization Towards Generalizable Re-ID

Masato Tamura
masato.tamura.sf@hitachi.com
Tomoaki Yoshinaga
tomoaki.yoshinaga.xc@hitachi.com
Lumada Data Science Lab.
Hitachi Ltd.
Tokyo, Japan

Abstract

In recent years, the performance of person re-identification has been dramatically improved by virtue of sophisticated training methods. However, most of the existing methods are based on the assumption that the statistics of a target domain can be utilized during training. This inevitably introduces huge costs for data collection each time a person re-identification system is deployed, which hinders the applicability to real-world scenarios. To mitigate this issue, we expand upon the concept of domain generalization. Typical person re-identification datasets are composed of a large amount of identities. However, examples for each identity are rather scarce. It is widely known that if examples are highly biased, over-fitting is likely to occur and degrade the performance. To alleviate this problem, we propose a novel soft-label regularization method that combines an expert feature extractor with a beginner classifier for generating soft labels. From a representation learning perspective, a convolutional neural network-based feature extractor is thought to prioritize common patterns. Therefore, the subsequent classifier typically fits common examples first, followed by rare ones. On the basis of this observation, we force the beginner classifier to remain uncertain towards rare examples by means of periodic initialization. Accordingly, the beginner classifier assigns highly confident labels to common examples and ambiguous labels to rare ones, thus enabling soft labels to mitigate over-fitting to biased examples (*e.g.*, highly occluded ones). Extensive analysis shows that our method successfully assigns ambiguous labels to biased examples and thus increases the rank-1 accuracy by 3.4 %, 1.6 %, 0.9 %, and 5.2 % on the VIPeR, PRID, GRID, and i-LIDS datasets, respectively. The source codes are available at https://github.com/hitachi-rd-cv/bcar.

1 Introduction

In recent years, person re-identification (Re-ID) has attracted much attention for its diverse range of real-world applications such as surveillance and marketing. The basic premise of such applications is to collect exact trajectories of individual pedestrians. Therefore, accurate Re-ID is crucial. Deep convolutional neural network (CNN)-based methods are usually applied for this and have demonstrated considerable improvement over the years [2, 10, 12, 14, 16, 18, 25, 26, 30]. However, drastic appearance changes caused by variations in illumination, viewpoints, poses, and occlusions remain a long-standing technical obstacle in conventional methods, which has spurred enduring interest in the topic of Re-ID.

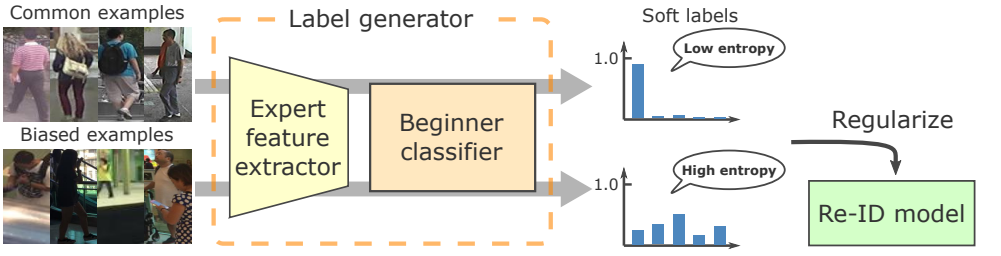


Figure 1: Overview of proposed soft-label regularization. Soft labels are generated by a model with a trained feature extractor and a classifier trained at the early stage of learning. This beginner classifier enables the generator to assign low-entropy labels to common examples and high-entropy labels to biased ones. Using these labels as regularization mitigates the performance degradation caused by biased examples.

Most existing approaches are based on the impractical assumption that the statistics of a target domain can be utilized during training. These approaches can be roughly divided into two categories: supervised training and unsupervised domain adaptation (UDA). In supervised training, all data are sampled from the same domain and labeled for identifying individuals. Therefore, Re-ID models do not suffer from severe domain shifts and thus can achieve promising results (*e.g.*, rank-1 accuracy of over 90 %) [16]. However, this approach introduces enormous costs for data collection and annotation each time a Re-ID system is deployed, which renders supervised training unsuitable for practical use. In UDA, training data and testing data are sampled from different domains, and testing data do not need to be annotated for training, which enables efficient utilization of existing large-scale datasets [2, 14, 25]. However, despite the reduction of costs for data annotation at the initial phase, data collection still incurs costs at each deployment.

The current best practice for real-world scenarios is domain generalization (DG). In DG, the statistics of a target domain remain unknown during training. Accordingly, the capability of acquiring domain-invariant knowledge is crucial for achieving high performance. For this purpose, two DG approaches have been proposed [10, 18], in which the conventional image classification architecture is modified to acquire the domain invariance. The modified models have outperformed existing supervised training and UDA methods on four publicly available benchmark datasets, which demonstrates the superiority of DG.

To further enhance the generalization performance of existing methods, we propose a beginner classifier as regularization (BCaR), a novel soft-label regularization method. Large-scale Re-ID datasets are generally composed of over thousands of identities. However, examples for each identity are rather scarce: sometimes as few as just two per identity. This is extremely small for training CNNs since their over-parameterized nature renders them prone to be over-fitted to a particular example. Various techniques such as weight decay, batch normalization [9], dropout [19], and label smoothing [21] have been proposed to prevent over-fitting, but as these techniques do not consider the dataset statistics, the performance improvement is limited. In contrast to these techniques, BCaR mitigates the impact of identities that have a few biased examples by utilizing soft labels. As Fig. 1 shows, soft labels are generated by an auxiliary generator model composed of a trained feature extractor and a classifier trained at the early stage of learning. This beginner classifier assigns low-entropy (*i.e.*, high-confident) labels to common examples and high-entropy (*i.e.*, ambiguous) labels

to rare ones because it is forced to remain uncertain towards rare examples due to the periodic initialization during generator training. This results in generating ambiguous labels for biased examples such as those of highly occluded pedestrians or of low brightness. By using these soft labels as regularization for Re-ID training, the impact of biased examples can be mitigated.

To summarize, our contributions are three-fold: (1) We propose BCaR, which utilizes soft labels for regularizing Re-ID models, as a remedy for over-fitting. BCaR enhances the generalization performance of existing models without incurring additional computational costs at test time. (2) We demonstrate state-of-the-art performances on Re-ID benchmarks using both MobileNet and ResNet backbones. (3) We conduct extensive experiments and show that our method improves performance even in within-dataset settings.

2 Related work

2.1 Generalized person re-identification

There have been a few prior studies that examined generalization performance [10, 18]. In [18], Song *et al*. proposed a model based on meta-learning [22] called Domain-Invariant Mapping Network (DIMN). In contrast to the common approach that utilizes feature distances for matching scores, DIMN generates classifier weights from gallery images and takes the dot product between the weights and probe image features to calculate matching scores. This meta-learning pipeline enables the model to be domain-invariant. However, its complicated learning procedure compounds the difficulties of optimization, and the weight generation during test time slows down the inference speed. Considering these drawbacks, an approach called DualNorm was proposed by Jia *et al*. [10]. It regards style and content variations as the cause of domain bias and suppresses them by inserting instance normalization [24] and batch normalization (BN) [9] at specific positions. They showed that the normalization successfully eliminates domain bias and improves the performance. Since our method can be applied as regularization, it can easily be combined with DualNorm to further enhance the performance without incurring additional computational costs at test time.

2.2 Soft-label regularization

Soft-label regularization is one of the most frequently used methods to alleviate over-fitting by imposing various distributed probabilities on each class. Label smoothing [21] is a simple soft-label approach that uniformly redistributes the probability of a ground-truth class to those of other classes. Although this approach is known to boost generalization performance, uniform redistribution is likely to be inconsistent. For comparatively consistent soft labels, learning-based label generation approaches have been proposed. Knowledge distillation [7] transfers knowledge from a high-capacity teacher model to a compact student model through soft labels. On the basis of knowledge distillation, Furlanello *et al*. proposed Born-Again Networks (BAN) [4] that utilize student and teacher models with identical structures. BAN shows that iterative training of a model with soft labels generated by a previously trained model improves performance. Recently, Tian *et al*. [23] proposed a method called network as regularization (NaR). As with BAN, NaR uses student and teacher models that have an identical structure, but differs in that both models are trained simultaneously using hard

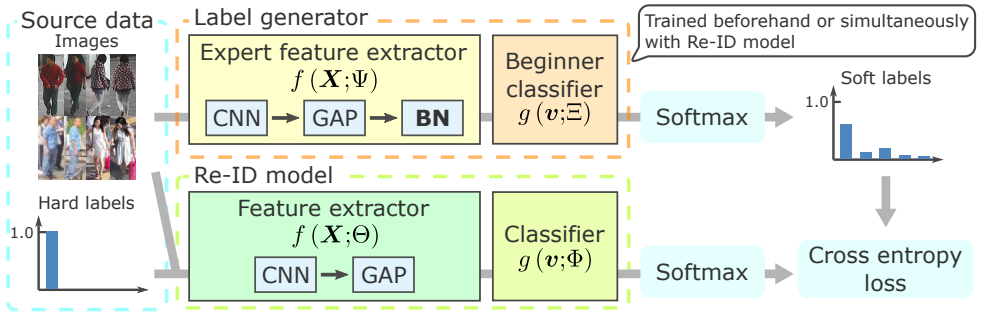


Figure 2: The proposed method. During Re-ID training, input images are forwarded to an auxiliary label generator for generating soft labels that regularize a Re-ID model. To generate effective soft labels, the classifier $g(\boldsymbol{v};\Xi)$ is forced to be in an early stage of learning by periodically initializing the parameter Ξ during generator training. A BN layer is inserted at the last part of the feature extractor $f(\boldsymbol{X};\Psi)$ for stable convergence after the initialization.

labels and dynamically generated soft labels. Although our proposed approach resembles NaR, the experiment in Sec. 4.6 shows that BCaR has a better performance.

3 Proposed method

3.1 Baseline

As a baseline method, we first introduce a naive deep learning approach called aggregation (AGG). Suppose we have K source domains $\mathcal{D} = \{D_k\}_{k=1}^{K}$, where $D_k = \left\{\left(\boldsymbol{X}^{(k)}, y^{(k)}\right)\right\}$, $\boldsymbol{X}^{(k)}$ is an image of an identity, and $y^{(k)}$ is a corresponding label. Here, we assume that there are N_k identities in the k-th domain D_k, and they are not overlapped between domains, *i.e.*, $\left\{y^{(k_1)}\right\} \cap \left\{y^{(k_2)}\right\} = \emptyset$ for arbitrary k_1 and k_2 $(1 \leq k_1 < k_2 \leq K)$. For training, all the source domains are combined into a single domain, hence the training dataset is $D = \{(\boldsymbol{X}, y)\} = \bigcup_{k=1}^{K} D_k$, and the number of identities is $N = \Sigma_{k=1}^{K} N_k$. During training, y is converted to a hard label $\boldsymbol{y}_{hard}$ in one-hot encoding, *i.e.*, $|\boldsymbol{y}_{hard}|_1 = 1$ and $\boldsymbol{y}_{hard} \in \{0,1\}^N$. Utilizing this label, a model is trained to minimize the following loss:

$$L_{AGG} = L(\boldsymbol{X}, \boldsymbol{y}_{hard}) = l^{(CE)}\left(g\left(f\left(\boldsymbol{X}\right)\right), \boldsymbol{y}_{hard}\right), \tag{1}$$

where $f(\cdot)$ is a feature extractor, $g(\cdot)$ is a classifier, and $l^{(CE)}(\cdot,\cdot)$ is the cross-entropy loss function. Once the loss converges, Re-ID is conducted using the cosine similarities of $f(\boldsymbol{X})$.

3.2 Training procedure

Figure 2 shows the proposed Re-ID training method. Hereinafter we denote a feature extractor in a Re-ID model as $f(\boldsymbol{X};\Theta)$, a classifier in that model as $g(\boldsymbol{v};\Phi)$, a feature extractor in a label generator model as $f(\boldsymbol{X};\Psi)$, and a classifier in that model as $g(\boldsymbol{v};\Xi)$, where Θ, Φ, Ψ, Ξ are the parameters of each component.

In the proposed training, an auxiliary label generator is used to generate soft labels. This generator is trained beforehand or simultaneously with the Re-ID model. As Algorithm 1

Algorithm 1: Training procedure for label generator

Data: Training dataset D
Parameter: LR η for $f(\boldsymbol{X};\Psi)$, LR η' for $g(\boldsymbol{v};\Xi)$, training epochs T, init interval τ

Initialize Ψ and Ξ
for $i \leftarrow 1$ **to** T **do**
  while *epoch not completed* **do**
    Create a mini-batch by randomly sampling image-label pairs from D
    Calculate the loss (Eq. 1)
    Update Ψ and Ξ with η and η', respectively
  if $\mathrm{mod}(i,\tau) = 0$ **then**
    Initialize Ξ

Algorithm 2: Training procedure for Re-ID model

Data: Training dataset D
Parameter: LR η for $f(\boldsymbol{X};\Theta)$ and $g(\boldsymbol{v};\Phi)$, training epochs T, regularization adj. α

Initialize Θ and Φ
for $i \leftarrow 1$ **to** T **do**
  while *epoch not completed* **do**
    Create a mini-batch by randomly sampling image-label pairs from D
    Generate soft labels (Eq. 2)
    Calculate the loss (Eq. 3)
    Update Θ and Φ with η

shows, given a training dataset D, the generator is trained using the loss Eq. 1 and stochastic gradient-based optimization. This is a quite common Re-ID training paradigm, but the parameter Ξ is initialized every few epochs for forcing the classifier $g(\boldsymbol{v};\Xi)$ to be in an early stage of learning. To control the upper bound of the capability of the classifier $g(\boldsymbol{v};\Xi)$ for fitting training data, the initialization interval τ and the learning rate (LR) η' are used as hyper-parameters. In the case that the generator is trained simultaneously with the Re-ID model, the generator is cloned before the initialization and the cloned model is used to generate soft labels in the next τ epochs after cloning.

The training procedure for a Re-ID model is shown in Algorithm 2. In contrast to the generator training, the Re-ID model is trained using both hard and soft labels. The soft labels are generated by the generator as follows:

$$\boldsymbol{y}_{soft} = \mathrm{Softmax}\left(g\left(f\left(\boldsymbol{X};\Psi\right);\Xi\right)\right). \tag{2}$$

Using those labels, the loss for the training is derived as

$$L_{BCaR} = \alpha L_{AGG} + (1-\alpha) L\left(\boldsymbol{X}, \boldsymbol{y}_{soft}\right), \quad 0 \leq \alpha \leq 1, \tag{3}$$

where α is a hyper-parameter that adjusts the magnitude of the soft-label regularization. Note that the loss is not back-propagated to the generator to keep it free from being affected by the Re-ID training.

3.3 Label generation network design

Since the parameter of the classifier in the label generator Ξ is periodically initialized during training, stable convergence is important. For this reason, a BN layer is inserted after global average pooling (GAP) in the feature extractor $f(\boldsymbol{X};\Psi)$. This design is similar to BNNeck [10, 16] but differs in that we insert a BN layer for the purpose of stabilizing convergence after the initialization of the parameter Ξ. The significance of the proposed method is further supported by the observation from the ablation study in Sec. 4.3 that combining BN with the initialization is more powerful than solely applying BN.

The classifier $g(\boldsymbol{v};\Xi)$ is the linear combination of an input variable and the parameter Ξ. Therefore, the gradient with respect to the parameter Ψ is derived as

$$\frac{\partial L}{\partial \Psi} = \left(\frac{\partial L}{\partial g}\right)\left(\frac{\partial g}{\partial \boldsymbol{v}}\right)\left(\frac{\partial \boldsymbol{v}}{\partial \Psi}\right) = \left(\frac{\partial L}{\partial g}\right)\Xi\left(\frac{\partial \boldsymbol{v}}{\partial \Psi}\right). \tag{4}$$

A concern arises that the layers in the feature extractor $f(\boldsymbol{X};\Psi)$ may be subject to the effect of the periodic initialization. Concretely, large initial values are likely to damage the feature extractor $f(\boldsymbol{X};\Psi)$. To mitigate this vulnerability, we initialize the parameter Ξ using a normal distribution with expected value 0 and standard deviation 0.001.

4 Experiments

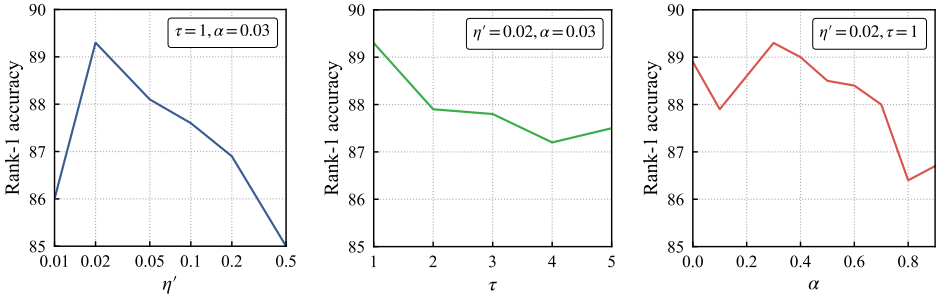
4.1 Datasets and evaluation settings

To evaluate our method, we follow the setting described in the papers of Song *et al*. [18] and Jia *et al*. [10]. In this setting, CUHK02 [11], CUHK03 [12], Duke MTMC [30], Market1501 [28], and PersonSearch [26] are combined into a single training dataset that has 121,765 images belonging to 18,530 identities. For evaluation, VIPeR [5], PRID [8], GRID [15], and i-LIDS [29] are used. Probe/gallery identities are randomly sampled from the overall identities in VIPeR, PRID, GRID, and i-LIDS in accordance with the number: 316/316, 100/649, 125/900, and 60/60, respectively. The evaluation is conducted in a single-shot manner on those sampled identities. For each dataset, we evaluate ten probe/gallery splits and report the average results.

4.2 Implementation details

We follow the paper of Jia *et al*. [10] and build our method by referring to its publicly available source code[1], within which MobileNetV2 [17] serves as the backbone network. Basically, the generator and the Re-ID model possess an identical structure, but in the case of the Re-ID model without a BN layer inserted after GAP (as aforementioned in Sec. 3.3), we solely insert a BN layer into the generator. The networks are trained from scratch for $T = 150$ epochs using stochastic gradient descent with the Nesterov momentum set to 0.9. The learning rate η is set to 0.01 and decayed by 0.1 after 100 epochs. Mini-batch size is set to 64, and all images in a mini-batch are resized to 256×128. To prevent over-fitting, weight decay of 0.0005, random horizontal flipping, and random cropping are used for all networks, whereas dropout [19] with the rate of 0.5 is used only for vanilla MobileNetV2 in the Re-ID model. For test-time data augmentation, horizontal flipping is used. To determine

[1] https://github.com/BJTUJia/person_reID_DualNorm



(a) (b) (c)

Figure 3: Rank-1 accuracy of a validation set. In (a), (b) and (c), the hyper-parameters η', τ, and α are changed respectively.

Table 1: Effect of each component. Rank-1 accuracy is shown.

Soft-label regularization	BN in generator	Periodic initialization	VIPeR	PRID	GRID	i-LIDS
			42.1	26.1	28.6	66.3
✓			42.4	25.1	30.0	65.5
✓	✓		41.2	35.5	30.6	64.8
✓		✓	37.3	11.0	22.1	56.5
✓	✓	✓	**50.4**	**37.1**	**31.9**	**68.7**

Table 2: Comparison of the generator training timing. Rank-1 accuracy is shown.

	VIPeR	PRID	GRID	i-LIDS
Before training the Re-ID model	44.3	30.5	31.5	67.2
Simultaneous training with Re-ID model	**50.4**	**37.1**	**31.9**	**68.7**

hyper-parameters η', τ, and α, we randomly select 2,000 identities from the training dataset and use them as a validation set. The validation results are shown in Fig. 3. In accordance with the results, we set $\eta' = 0.02$, $\tau = 1$, and $\alpha = 0.3$. Note that the optimizer's momentum of the parameter Ξ is also reset when the parameter is initialized, and the learning rate η' is not decayed so as to preserve the upper bound of the capability of the classifier $g(\boldsymbol{v};\Xi)$ for fitting training data.

4.3 Ablation study

To investigate the effect of each essential component (*i.e.*, the soft-label regularization, BN in the generator, and periodic initialization) in the proposed BCaR, we introduce four ablated variants based on the AGG baseline [10]. Table 1 shows the comparison results. We can see that solely applying the soft-label regularization, which is almost identical to NaR [23], yielded improvement only on VIPeR and GRID, and coupling BN in the generator with the soft-label regularization yielded improvement only on PRID and GRID. These results indicate that universal improvement across all datasets cannot be obtained by just training the Re-ID model with an auxiliary label generator network. Additionally, we found that

Table 3: Comparison results against baselines. (R: Rank, S: Supervised training, U: UDA, DG: Domain generalization with the MobileNetV2 backbone, –: No report)

		VIPeR		PRID		GRID		i-LIDS	
	Type	R-1	R-10	R-1	R-10	R-1	R-10	R-1	R-10
SpindleNet [27]	S	53.8	83.2	**67.0**	89.0	–	–	66.3	91.8
SSM [1]	S	53.7	**91.5**	–	–	27.2	61.2	–	–
JLML [13]	S	50.2	84.3	–	–	37.5	69.4	–	–
SSDAL [20]	U	37.9	75.6	20.1	55.7	19.1	45.8	–	–
TJAIDL [25]	U	38.5	–	34.8	–	–	–	–	–
MMFAN [14]	U	39.1	–	35.1	–	–	–	–	–
Synthesis [2]	U	43.0	–	43.0	–	–	–	56.5	–
AGG [10]	DG	42.1	–	27.2	–	28.6	–	66.3	–
DIMN [18]	DG	51.2	76.0	39.2	76.7	29.3	65.8	70.2	94.5
DualNorm [10]	DG	53.9	–	60.4	–	41.4	–	74.8	–
Vanilla net + BCaR	DG	50.4	77.0	37.1	68.4	31.9	61.6	68.7	94.3
DualNorm + BCaR	DG	**57.3**	83.3	62.0	**89.8**	**42.3**	**74.5**	**80.0**	**97.0**

Table 4: Comparison with the ResNet backbone. Rank-1 accuracy is shown.

Method	VIPeR	PRID	GRID	i-LIDS
AGG [10]	48.5	20.3	29.0	71.3
DualNorm [10]	59.4	69.6	43.7	78.2
DualNorm + BCaR	**65.8**	**70.2**	**52.8**	**81.3**

combining the soft-label regularization and periodic initialization degraded performance. This is because without BN, the training of the classifier in the generator takes long time, and it becomes difficult to determine the initialization timing for generating effective soft labels. By applying the periodic initialization tactics with BN in the generator, universal improvement across all datasets was demonstrated.

As described in Sec. 3.2, the generator can be trained before training the Re-ID model or simultaneously with it. To unravel the optimal training timing of the generator, we conducted a comparison with the generator trained beforehand and simultaneously with the Re-ID model. Table 2 shows the comparison result. The result showed that the simultaneous training yielded better performance. The reason for this is that the simultaneous training can mitigate the dependence of the input order of training data. Since the classifier in the generator is trained for only a epoch, the output soft labels are highly dependent on the input order. However, in the simultaneous training, the classifier is updated at every epoch and outputs different soft labels. Therefore, the impact of the dependency to the Re-ID model is mitigated, and the effectiveness of the soft labels is boosted. Considering this performance advantage, the simultaneous setting is used in the following experiments.

4.4 Comparison against baselines

To evaluate the performance of BCaR, we conducted a comparison with previously proposed baselines. In addition to the results of DG methods, we provide those of supervised training

Table 5: Within-dataset results on Market1501 and Duke MTMC. (R: Rank)

Dataset	Method	R-1	R-5	R-10	mAP
Market1501	Vanilla net (MobileNetV2) [10]	77.2	89.9	93.8	53.9
	DualNorm (MobileNetV2) [10]	82.6	91.7	95.3	57.2
	DualNorm (MobileNetV2) + BCaR	**87.6**	**95.5**	**97.1**	**65.7**
Duke MTMC	Vanilla net (MobileNetV2) [10]	65.0	79.8	84.1	44.1
	DualNorm (MobileNetV2) [10]	71.2	82.5	86.3	48.3
	DualNorm (MobileNetV2) + BCaR	**78.6**	**86.5**	**89.6**	**55.2**

methods and UDA methods for reference. As shown in Table 3, BCaR had a competitive or even better performance than the supervised training methods and UDA methods, indicating that combining a large amount of existing training data with the proposed regularization can achieve state-of-the-art performance without the need for cumbersome data collection. Comparing within DG, BCaR improved the rank-1 accuracy by 8.3 %, 9.9 %, 3.3 %, 2.4 % for vanilla MobileNetV2 and by 3.4 %, 1.6 %, 0.9 %, and 5.2 % for DualNorm on VIPeR, PRID, GRID, and i-LIDS, respectively. This shows that BCaR can further enhance the performance of the outstanding baselines without adding any computational burden at test time.

We also evaluate BCaR with the ResNet50 [6] backbone. Following the paper of Jia *et al*. [10], the model is initialized with an ImageNet [3] pre-trained model and trained for 70 epochs. The learning rate for the feature extractors $f(\boldsymbol{X};\Theta)$ and $f(\boldsymbol{X};\Psi)$ is set to 0.005, and that for the classifier $g(\boldsymbol{v};\Phi)$ is set to 0.05. Those rates are decayed by 0.1 after 40 epochs. The hyper-parameters are set to the values of the evaluation with the MobileNetV2 backbone. Table 4 shows the comparison results. Again, BCaR improved the generalization performance, thus demonstrating the model-independent effectiveness of BCaR.

4.5 Within-dataset evaluation

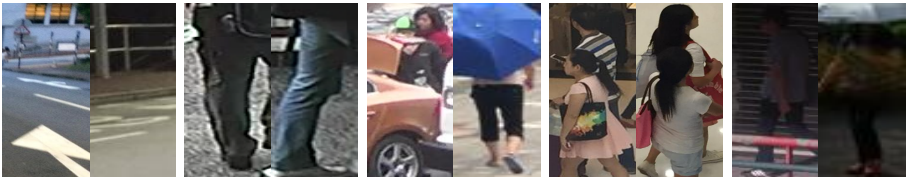
Next, we investigate invariant effectiveness of BCaR under a supervised setting. In this evaluation, two large-scale datasets, Market1501 [28] and Duke MTMC [30], are used. We follow the original train/test splits and evaluation protocols. The implementation remains the same as that described in Sec. 4.2 except that we set the mini-batch size to 16, since the reported results in [10] are acquired with that mini-batch size. As shown in Table 5, BCaR improved the rank-1 accuracy by 5.0 % and 7.4 % for Market1501 and Duke MTMC, respectively, suggesting that biased examples degrade the performance even within the same domain, and BCaR successfully mitigates the impact of those examples.

4.6 Analysis on soft-label regularization

To further examine our proposed soft-label regularization, comparison with other soft-label regularization methods is conducted. As criteria for evaluation, the mean and standard deviation of the entropy are calculated for the teacher soft labels of each method. DualNorm with the MobileNetV2 backbone serves as the model for this comparison. In Table 6, the results of the second generation of BAN [4], the third generation of BAN, NaR [23], and our BCaR are shown. Note that the entropy of BCaR is averaged over the last 20 epochs, as the periodic initialization makes soft labels highly sensitive to the order in which training data are input. We can see that BAN-3 yielded the entropy of higher mean and standard deviation

Table 6: Analysis on various soft-label regularization methods. Rank-1 accuracy is shown.

Method	Mean of label entropy	Standard deviation of label entropy	VIPeR	PRID	GRID	i-LIDS
BAN-2 [4]	1.23	0.78	46.6	54.6	38.4	72.7
BAN-3 [4]	2.27	1.19	47.9	61.5	38.2	72.5
NaR [23]	1.33	0.75	46.9	58.1	38.4	71.2
BCaR	1.32	1.46	**57.3**	**62.0**	**42.3**	**80.0**



(a) (b) (c) (d) (e)

Figure 4: Example of images with soft labels of high entropy. (a) No person. (b) Only part of a person. (c) Highly occluded person. (d) Behind another person. (e) Low brightness.

than BAN-2 and NaR, outperforming them by a narrow margin. This indicates that BAN-3 imposes a stronger yet example-dependent regularization, which works slightly better than the other two. Compared with BAN-3, BCaR yielded lower mean yet higher standard deviation of the entropy, revealing that our regularization is highly example-dependent. Figure 4 shows sample images with soft labels of high entropy. These images are of poor quality, indicating that ambiguous labels were successfully assigned to biased examples, and that our regularization works well for mitigating the impact of those examples.

5 Conclusion

In this paper, we have proposed a novel beginner classifier as soft-label regularization to boost the generalization performance of Re-ID. The beginner classifier is in the soft label generator and is forced to remain uncertain towards biased examples due to periodic initialization. Our empirical exploration demonstrates that the proposed classifier assigns soft labels of high entropy to the biased examples, thus mitigating their impact on the model's generalization performance. Cross-domain evaluations conducted on four public benchmark datasets demonstrate a state-of-the-art performance. Additionally, we found that our method can further improve the performance of existing Re-ID models without increasing the computational burden at test time.

References

[1] Song Bai, Xiang Bai, and Qi Tian. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, July 2017.

[2] Slawomir Bak, Peter Carr, and Jean-Franois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *ECCV*, September 2018.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei Fei Li. ImageNet: A large-scale hierarchical image database. In *CVPR*, June 2009.

[4] Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born-again neural networks. In *ICML*, 2018.

[5] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, October 2008.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, June 2016.

[7] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Workshop*, 2015.

[8] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, May 2011.

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, July 2015.

[10] Jieru Jia, Qiuqi Ruan, and Timothy M. Hospedales. Frustratingly easy person re-identification: Generalizing person re-id in practice. In *BMVC*, September 2019.

[11] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *CVPR*, June 2013.

[12] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. DeepReID: Deep filter pairing neural network for person re-identification. In *CVPR*, June 2014.

[13] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. In *IJCAI*, August 2017.

[14] Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex Chichung Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In *BMVC*, September 2018.

[15] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera activity correlation analysis. In *CVPR*, June 2009.

[16] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshops*, June 2019.

[17] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *CVPR*, June 2018.

[18] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *CVPR*, June 2019.

[19] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, January 2014.

[20] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi" Tian. Deep attributes driven multi-camera person re-identification. In *ECCV*, September 2016.

[21] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, June 2016.

[22] Sebastian Thrun and Lorien Pratt, editors. *Learning to Learn*. Kluwer Academic Publishers, 1998.

[23] Kai Tian, Yi Xu, Jihong Guan, and Shuigeng Zhou. Network as regularization for training deep neural networks: Framework, model and performance. In *AAAI*, Februaly 2020.

[24] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[25] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*, June 2018.

[26] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*, July 2017.

[27] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle Net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, July 2017.

[28] Liang Zheng, Liyue Shen, Lili Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, December 2015.

[29] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *BMVC*, September 2009.

[30] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, October 2017.