Debiased Dual-Invariant Defense for Adversarially Robust Person Re-Identification

Yuhang Zhou[1], Yanxiang Zhao[1], Zhongyun Hua[1*], Zhipu Liu[2], Zhaoquan Gu[1,3], Qing Liao[1,3] , Leo Yu Zhang[4]

[1] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China
[2] School of Computer Science and Engineering, Chongqing University of Technology, China
[3] Peng Cheng Laboratory, China
[4] School of Information and Communication Technology, Griffith University, Australia
23B951015@stu.hit.edu.cn, 24S151088@stu.hit.edu.cn, huazhongyun@hit.edu.cn, zpliu@cqut.edu.cn, guzhaoquan@hit.edu.cn, liaoqing@hit.edu.cn, leo.zhang@griffith.edu.au

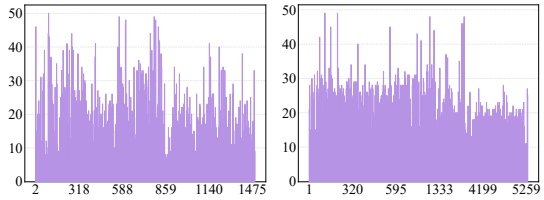
Abstract

Person re-identification (ReID) is a fundamental task in many real-world applications such as pedestrian trajectory tracking. However, advanced deep learning-based ReID models are highly susceptible to adversarial attacks, where imperceptible perturbations to pedestrian images can cause entirely incorrect predictions, posing significant security threats. Although numerous adversarial defense strategies have been proposed for classification tasks, their extension to metric learning tasks such as person ReID remains relatively unexplored. Moreover, the several existing defenses for person ReID fail to address the inherent unique challenges of adversarially robust ReID. In this paper, we systematically identify the challenges of adversarial defense in person ReID into two key issues: model bias and composite generalization requirements. To address them, we propose a debiased dual-invariant defense framework composed of two main phases. In the data balancing phase, we mitigate model bias using a diffusion-model-based data resampling strategy that promotes fairness and diversity in training data. In the bi-adversarial self-meta defense phase, we introduce a novel metric adversarial training approach incorporating farthest negative extension softening to overcome the robustness degradation caused by the absence of classifier. Additionally, we introduce an adversarially-enhanced self-meta mechanism to achieve dual-generalization for both unseen identities and unseen attack types. Experiments demonstrate that our method significantly outperforms existing state-of-the-art defenses.

Code — https://github.com/zchuanqi/DDDefense-ReID
Extended version — https://arxiv.org/abs/2511.09933

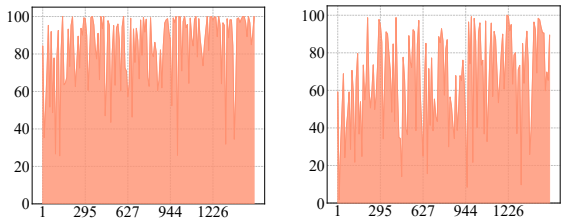
Introduction

Person re-identification (ReID) (Zhou et al. 2023; Ye et al. 2021; Wei et al. 2021) aims to retrieve a specific identity (ID) of interest from an image gallery. This capability enables practical applications such as tracking a suspect's movement trajectory from a single photo. The rapid development of deep learning provides high-performance ReID solutions.



(a) Market-1501 (b) DukeMTMC

Figure 1: Statistics of sample counts for each ID.



(a) The mAP of each ID for vanilla-trained model (avg.=84.13 and std.=18.78).

(b) The mAP of each ID for adversarially-trained model (avg.=66.10 and std.=23.40).

Figure 2: Biased model accuracy.

However, deep neural networks are vulnerable to adversarial attacks (Goodfellow, Shlens, and Szegedy 2014; Madry et al. 2018; Croce and Hein 2020), where the addition of human-imperceptible perturbations to input data can cause the model to produce completely incorrect predictions. This vulnerability also extends to person ReID systems, raising even more serious and realistic security threats. For example, malicious actors may manipulate inputs to evade detection or tracking by ReID model-based surveillance systems.

Several adversarial attack methods targeting person ReID have been proposed (Bai et al. 2020; Wang et al. 2019; Zheng et al. 2023; Bouniot, Audigier, and Loesch 2020). Early works focused on metric-based white-box attacks (Bai et al. 2020; Bouniot, Audigier, and Loesch 2020), which revealed the vulnerability of ReID models when full access to

*Corresponding author

Figure 3: Partial adversarial examples from the original dataset illustrating the challenge of homogenization.

Models	Defense	clean	PGD
ResNet18	AT	87.273	54.045
	AT → fine-tuned(only classifier)	88.113	50.752
ResNet50	AT	88.786	53.076
	AT → fine-tuned(only classifier)	89.161	50.218

Table 1: Validation of the hypothesis that "partial robustness knowledge is distributed on the classifier".

the model is available. Subsequently, black-box attack methods are developed, enabling adversarial manipulation without direct access to the target model (Liu et al. 2023; Zhang et al. 2020; Tsipras et al. 2018; Zhang et al. 2019; Raghunathan et al. 2019). In contrast to the research on adversarial attacks, defense strategies for person ReID remain relatively scarce. Bai et al. (2020) propose an offline adversarial training method based on adversarial metric samples. Bian et al. (2025) leverage virtual data to enhance the transferability of adversarial robustness. More recently, Wei et al. (2024) develop a dynamic attack budget strategy to enhance defense effectiveness. However, these existing methods fail to consider the unique characteristics of person ReID, resulting in robustness degradation. In the supplementary material, we provide a detailed related works, including *adversarial attacks, defenses*, and *person ReID*.

In this work, we systematically re-examine the unique challenges of adversarial defense in person ReID and identify two key factors that contribute to the limitations of existing defense approaches:

Model bias. We attribute the model bias to two main issues. *(i)* ReID datasets exhibit significant variation in inter-ID samples number, since the sample number for each ID depends on the frequency of its appearance under the camera. To demonstrate it, we report the per-identity sample distributions in two widely used datasets: DukeMTMC (Zheng et al. 2015) and Market-1501 (Zheng et al. 2016). The results are shown in Figure 1. This imbalance leads to biased and unfair feature representations, as evidenced by the inter-ID accuracy variance of models trained using vanilla training (cross-entropy + triplet loss) and standard adversarial training (Bouniot, Audigier, and Loesch 2020), shown in Figure 2. *(ii)* Datasets are typically extracted from video sequences, often resulting in high intra-ID redundancy and limited visual diversity, as illustrated in Figure 3. This homogeneity constrains the amount of useful information each identity provides during training, as visually similar frames contribute little to entropy or feature discrimination.

Composite generalization requirements. Compared to other common tasks (i.e., classification task), ReID task shows two unique generalization requirements. *(i)* Adversarial training inevitably allocates partial robustness to the classifier which is unusable for ReID during testing, resulting in a decrease in robustness. To verify this issue, we adopt standard adversarial training to obtain a robust model A on CIFAR10. Then, we freeze the feature encoder and fine-tune the classifier on clean samples, resulting in robust model B with a vanilla classifier. The evaluation results of models A and B, as shown in Table 1, indicate that the clean accuracy of the model after fine-tuning does not change largely, but robustness decreases significantly. This indicates that partial robustness knowledge is distributed on the classifier in adversarial training. Consequently, the feature encoder requires enhanced adversarial knowledge for generalized representation to reduce its dependency on the classifier's robustness. *(ii)* Adversarial defense for person ReID is an open-set task while the attack strategies are also too diverse to enumerate during training. Consequently, adversarially robust ReID requires dual-dimensional adversarial generalization for both unseen IDs and unseen attacks, which poses more challenge than close-set task on which nowadays mainstream defense methods focus.

To address the above challenges, we propose a novel composite adversarial defense framework comprising two key phases: a data balancing phase and a bi-adversarial self-meta defense phase. In the data balancing phase, we employ a diffusion-model-based augmentation strategy to optimize both inter-ID and intra-ID sample distributions. In the bi-adversarial self-meta defense phase, to address the robustness degradation caused by the absence of the classifier, we propose a novel metric adversarial training strategy based on farthest negative extension softening, which promotes robust representation learning by enhancing adversarial diversity. Additionally, we incorporate label softening to mitigate the overfitting commonly caused by hard-label supervision. To further support generalization across both unseen identities and attack types, we introduce an adversarially-enhanced self-meta defense to extract dual-invariant features: inter-ID invariant features and adversarial-clean invariant features. Here, adversarial metric training and adversarial learning together form the bi-adversarial framework. Our main contributions can be summarized as follows:

- We identify and empirically validate two fundamental challenges specific to adversarial defense in person ReID: model bias and the composite generalization requirement across both unseen identities and attack types.
- To address these challenges, we propose a novel composite defense framework for person ReID, integrating data balancing and bi-adversarial self-meta defense.
- Experiments demonstrate that our method achieves state-of-the-art performance. Ablation studies confirm the contribution of each component, while transfer and interpretability evaluations highlight the framework's generalization capability and practical applicability.

Methods

Preliminaries

Let E, parameterized by θ_E, denote the feature encoder, and H, parameterized by θ_H, denote the classifier. Together, they

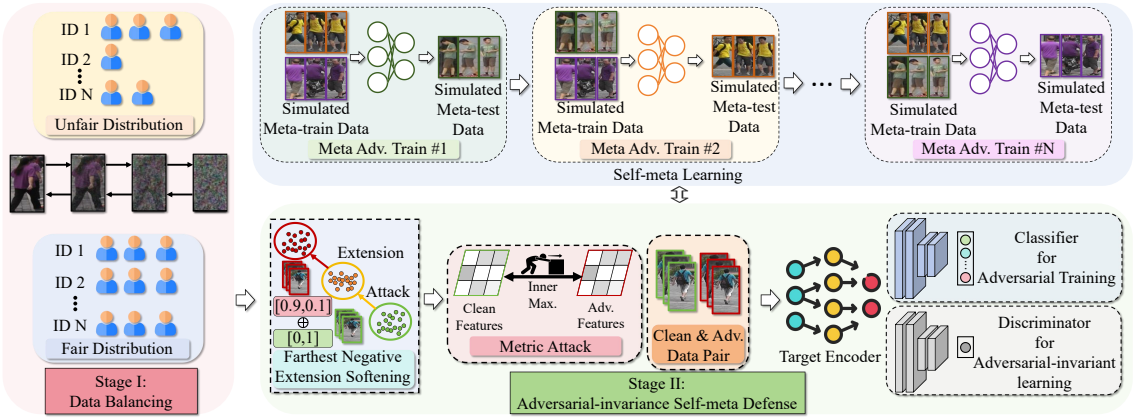


Figure 4: Overview of our proposed method, which consists of data balancing and bi-adversarial self-meta defense.

form the complete model $G = H(E(\cdot))$, with parameters $\theta_G = \theta_E \cup \theta_H$. During training, both E and H are jointly optimized. However, during testing, only the feature encoder E is used to extract features for retrieval. We define the target dataset distribution as $(x, y) \sim \mathcal{D}$, where x represents the input samples and y represents the corresponding identity labels. Our method consists of two primary stages: data balancing and bi-adversarial self-meta defense.

Data Balancing

In this stage, we aim to address the fairness and diversity issues caused by data imbalance through a targeted data balancing strategy.

Data balancing via diffusion model. To mitigate model bias caused by inter-ID data imbalance and limited intra-ID data diversity, we introduce a diffusion-model-based data balancing approach. Specifically, we train corresponding diffusion models on ReID datasets under an ID-conditional setting. For IDs with a small number of samples, we synthesize pseudo samples until a predefined sample threshold is met. The inter-ID data balancing process is formulated as follows:

$$\mathcal{D} \leftarrow \mathcal{D} \cup \left\{ x_i^{\text{pseudo},j} \mid i \in \mathcal{I},\ n_i < \delta_1,\ j = 1, \ldots, \delta_1 - n_i \right\}, \tag{1}$$

where x_i^{pseudo} denotes samples generated by the diffusion model with ID i, the set $\mathcal{I}$ denotes all IDs from the distribution $\mathcal{D}$, n_i is the initial sample number of i and δ_1 is the threshold. Subsequently, IDs with a proportion of samples from any single camera exceeding a predefined threshold are deemed to lack diversity. For these IDs, we synthesize pseudo samples from other cameras. The intra-ID diversifying process is formulated as follows:

$$\mathcal{D}_{i'} \leftarrow \mathcal{D}_{i'} \cup \left\{ x_{i',c}^{\text{pseudo}} \mid c \in \mathcal{C} \backslash \{c_{i'}\} \right\}, c_{i'} = \arg\max_{c \in \mathcal{C}} \frac{n_{i',c}}{n_{i'}} \tag{2}$$

where $\mathcal{D}_{i'}$ is the subset of $\mathcal{D}$ corresponding to ID i', $x_{i',c}^{\text{pseudo}}$ denotes pseudo samples with ID i' and camera c, the set $\mathcal{C}$ denotes all cameras of $\mathcal{D}$, $n_{i',c}$ is the sample number of i' from c. i' satisfies $i' \in \mathcal{I}$ and $\max_{c \in \mathcal{C}} \frac{n_{i',c}}{n_{i'}} > \delta_2$, and δ_2 is the predefined proportion threshold. The number of $x_{i',c}^{\text{pseudo}}$ is the initial mean value of each camera of $\mathcal{D}_{i'}$. In this implementation, we adopt the EDM framework (Karras et al. 2022).

Bi-adversarial Self-meta Defense

In this stage, we propose a novel metric adversarial training based on a farthest negative extension softening method to address the robustness decreasing caused by the absence of the classifier with better attack diversity, and introduce an adversarially-enhanced self-meta defense to learning dual-invariant features: adversarial invariant features shared between clean samples and adversarial samples, and generalization invariant features shared between seen IDs and unseen IDs. Bi-adversarial framework consists of adversarial metric training and adversarial learning.

Adversarial metric training based on farthest negative extension softening. The classifier H is discarded during inference, where only features extracted by E are used for computing distance metric. Consequently, current adversarial training methods for ReID commonly use metric variants of PGD (Bouniot, Audigier, and Loesch 2020):

$$\begin{aligned} \hat{x}^{(0)} &= x + \eta, \\ \hat{x}^{(n+1)} &= \Psi_x^{\epsilon}\left(\hat{x}^{(n)} + \kappa \cdot \text{sign}\left(\frac{\partial \mathcal{L}_{\text{metric}}(\hat{x}^{(n)})}{\partial \hat{x}^{(n)}} \right) \right), \\ \mathcal{L}_{\text{metric}}(x) &= \sum_{x^p \in \mathcal{P} \cap \mathcal{A}} d\big(E(x), E(x^p)\big) \\ &\quad - \sum_{x^n \in \mathcal{N} \cap \mathcal{A}} d\big(E(x), E(x^n)\big), \end{aligned} \tag{3}$$

where x is the clean sample, $\hat{x}^{(n)}$ is the adversarial sample after the n-th iteration, η is the initial perturbation, ϵ is the perturbation budget and κ is the perturbation step size. Ψ_x^{ϵ} is the clip function, which ensure that $\left\| \hat{x}^{(n+1)} - x \right\|_{\infty} \leq \epsilon$. The set $\mathcal{A}$ denotes other accessible samples, typically the current batch during training or the query set during inference. The set $\mathcal{P}$ denotes the samples with the same ID as x and $\mathcal{N}$ denotes the samples belonging to the ID farthest from x. We use the euclidean distance as the distance metric $d(\cdot)$.

To address the robustness decreasing caused by the absence of the classifier, we introduce farthest negative ex-

Model	Dataset	Defense	Clean	FNA		SMA		IFGSM	
				8/255-16	10/255-16	8/255-16	10/255-16	8/255-16	10/255-16
ResNet50	Market	Origin	78.49/92.01	0.20/0.17	0.18/0.14	0.27/0.26	0.20/0.11	1.25/1.95	1.09/1.66
		BitSqueezing	77.46/91.54	0.22/0.18	0.20/0.12	0.31/0.24	0.25/0.12	1.03/1.40	0.95/1.16
		MedianSmoothning2D	**78.55/91.95**	0.68/0.89	0.57/0.80	3.57/4.84	3.29/3.86	3.77/6.41	3.33/5.70
		BS+MS	77.22/90.06	1.33/2.08	1.08/1.84	6.98/10.15	5.90/8.67	6.31/11.58	5.57/10.27
		AMD	76.85/90.97	0.30/0.36	0.26/0.27	0.45/0.39	0.42/0.33	1.41/2.35	1.17/2.02
		Adv_train	69.69/88.24	8.57/18.14	4.37/9.41	22.85/35.69	15.21/23.37	17.97/34.65	11.74/23.34
		DAS	69.79/88.39	12.70/24.85	7.25/14.52	32.14/49.05	24.37/38.69	22.33/39.79	15.90/30.93
		Ours	68.50/88.21	**31.99/55.17**	**24.80/45.93**	**50.13/72.60**	**45.96/67.34**	**37.61/62.02**	**31.20/54.22**
	Duke	Origin	68.83/83.80	0.10/0.00	0.09/0.00	0.30/0.18	0.24/0.13	0.89/1.17	0.79/1.03
		BitSqueezing	68.36/83.03	0.13/0.18	0.12/0.13	0.64/0.58	0.50/0.27	1.44/2.15	1.25/1.66
		MedianSmoothning	69.46/84.02	0.34/0.31	0.27/0.22	3.80/4.85	3.48/4.44	3.72/6.60	3.09/5.12
		BS+MS	68.04/83.21	0.58/0.63	0.44/0.36	6.36/8.62	5.07/6.87	5.45/8.98	4.58/7.50
		AMD	**69.56/84.02**	0.10/0.00	0.10/0.00	0.46/0.36	0.38/0.31	1.08/1.53	0.97/1.39
		Adv_train	57.29/75.31	7.47/13.64	3.86/7.68	18.46/29.44	13.28/20.65	16.64/29.94	12.07/22.31
		DAS	59.04/77.33	11.15/19.84	5.93/11.27	26.09/39.45	19.36/29.89	19.38/33.39	13.70/24.55
		Ours	55.29/75.81	**26.02/43.49**	**20.76/36.54**	**40.74/59.69**	**37.50/55.83**	**31.23/50.67**	**26.89/45.29**

Table 2: White-box robustness results on ResNet50.

Models	datasets	Modules	Clean	8/255-16			10/255-16		
				FNA	SMA	IFGSM	FNA	SMA	IFGSM
ResNet50	Market	Metric AT	67.20/88.00	28.38/52.26	45.38/68.74	33.97/58.52	22.02/42.96	40.56/63.18	28.57/52.23
		+Diffusion model	66.96/86.91	29.85/53.36	45.82/68.41	35.32/59.68	22.82/43.38	41.14/63.26	29.03/52.46
		+Adversarial learning	67.81/88.21	30.34/53.77	48.10/70.72	35.70/60.27	23.21/44.21	43.58/65.17	29.31/52.29
		+Self-meta learning	68.24/88.03	29.48/52.88	46.91/69.30	35.09/59.86	22.19/42.96	42.06/64.52	28.61/52.31
		+FNES	68.29/88.07	30.98/54.45	49.25/71.11	37.16/61.49	24.64/44.83	44.07/65.97	30.54/53.15
		All modules	**68.50/88.21**	**31.99/55.17**	**50.13/72.60**	**37.61/62.02**	**24.80/45.93**	**45.96/67.34**	**31.20/54.22**
	Duke	Metric AT	54.47/74.55	23.76/39.27	40.40/57.99	27.04/44.21	17.41/30.97	35.98/52.42	21.55/36.58
		+Diffusion model	55.15/74.87	24.82/42.48	39.44/57.14	30.10/49.31	18.67/35.45	35.50/53.46	24.93/43.67
		+Adversarial learning	54.55/75.13	25.38/42.03	38.67/57.71	30.75/49.55	19.72/34.86	36.08/54.67	25.97/42.91
		+Self-meta learning	54.88/74.96	25.12/43.02	38.65/58.57	30.11/49.69	19.53/35.02	35.87/54.30	25.16/43.60
		+FNES	54.81/75.40	25.20/42.75	40.18/58.14	30.23/49.53	19.82/35.21	36.89/54.11	25.60/43.57
		All modules	**55.29/75.81**	**26.02/43.49**	**40.74/59.69**	**31.23/50.67**	**20.76/36.54**	**37.50/55.83**	**26.89/45.29**

Table 3: Ablation analysis of our proposed model on ResNet50.

tension softening (FNES) for robust representation learning with better attack diversity. Specifically, after attacking, we apply linear scaling to the final adversarial perturbation and soften the label of adversarial sample on the farthest negative class as:

$$\begin{aligned} x^{\text{temp}} &= x + \gamma \cdot (\hat{x} - x)\,, \\ x^{\text{adv}} &= \omega x + (1-\omega) x^{\text{temp}}, \\ y^{\text{adv}} &= \omega \phi(y, \lambda_1) + (1-\omega)\tau\big(\phi(y, \lambda_2), \upsilon\big), \end{aligned} \tag{4}$$

where x^{temp} is the temporary adversarial sample after linear scaling, $\hat{x}$ is the adversarial sample before FNES, $\gamma \geq 1$ is a scaling factor, ω is a real value sampled from a uniform distribution $\mathcal{U}(a, b)$ with $a, b \in (0, 1)$ and x^{adv} is the final adversarial sample obtained by linearly mixing x and x^{temp}. ϕ is the label-smoothing function (Szegedy et al. 2016), assigning $\lambda \in (0, 1)$ to the true class and equally distributing $\frac{1-\lambda}{k-1}$ to the other classes in the case of k classes. τ is a label operation that redistributes a portion $\upsilon \in (0, \lambda_2)$ from the true class to the farthest negative class which $\mathcal{N}$ belongs to. In this manner, linearly scaling perturbation mitigates the high similarity among adversarial samples by overcoming fixed iteration directions in metric-based PGD and diversifying the attack budget. Furthermore, assigning a portion of the labels to farthest negative classes corresponds to the behavior of metric attacks where samples are pulled toward farthest negative classes. This label softening operation facilitates models' learning of robustness-related knowledge about farthest negative classes and alleviates the overfitting problem caused by hard-label based training.

The max-min optimization process for adversarial training based on FNES is formulated as follows:

$$\begin{aligned} &\max_{\hat{x}} \mathbb{E}_{\hat{x}} \left[\mathcal{L}_{\text{metric}}(\hat{x})\right], \\ &\min_{G} \mathbb{E}_{(x^{\text{adv}}, y^{\text{adv}})} \left[\mathcal{L}_{\text{cls}}\big(G(x^{\text{adv}}), y^{\text{adv}}\big) + \mathcal{L}_{\text{tri}}\big(E(x^{\text{adv}})\big)\right], \\ &\mathcal{L}_{\text{tri}}(f) = \left[d(f, f^+) - d(f, f^-) + m\right]_+, \end{aligned} \tag{5}$$

where $\hat{x}$ and x^{adv} respectively denote the adversarial samples before and after being processed by FNES, $\mathcal{L}_{\text{cls}}$ is the classification loss and $\mathcal{L}_{\text{tri}}$ is the triplet loss with margin value of m. f^+ and f^- respectively denote the sample features sharing the same ID as x and having a different ID from x.

Adversarially-enhanced learning. To learn adversarial-invariant features that are shared between clean samples and adversarial samples, we introduce a feature discriminator D parameterized by θ_D, which, together with the feature encoder E, forms an adversarial learning framework. Specifically, the discriminator D is trained to distinguish whether the features originate from x^{adv} or x, the loss of which can be formalized as:

$$\mathcal{L}_D = -\mathbb{E}_x[\log D(E(x))] - \mathbb{E}_{x^{\text{adv}}}\left[\log\left(1 - D\left(E\left(x^{\text{adv}}\right)\right)\right)\right]. \tag{6}$$

The objective of E is to maximally confuse the discriminator D, thereby preventing D from determining whether the output features originate from x or x^{adv}. This objective can be formalized as maximizing the loss of D:

$$\mathcal{L}_E = \mathbb{E}_x[\log D(E(x))] + \mathbb{E}_{x^{\text{adv}}}\left[\log\left(1 - D\left(E\left(x^{\text{adv}}\right)\right)\right)\right]. \tag{7}$$

Models	Datasets	Defense	Clean	8/255-16			10/255-16		
				FNA	SMA	IFGSM	FNA	SMA	IFGSM
ResNet50	Market	AT_PGD	64.24/85.48	26.18/47.86	36.32/58.22	27.86/50.36	19.25/38.39	30.72/50.53	21.26/41.57
		Metric AT	67.20/88.00	28.38/52.26	45.38/68.74	33.97/58.52	22.02/42.96	40.56/63.18	28.57/52.23
		Ours	**68.50/88.21**	**31.99/55.17**	**50.13/72.60**	**37.61/62.02**	**24.80/45.93**	**45.96/67.34**	**31.20/54.22**
	Duke	AT_PGD	53.60/73.70	22.34/38.59	33.40/50.72	26.48/44.03	16.69/29.59	29.16/45.47	21.33/35.61
		Metric AT	54.47/74.55	23.76/39.27	40.40/57.99	27.04/44.21	17.41/30.97	35.98/52.42	21.55/36.58
		Ours	**55.29/75.81**	**26.02/43.49**	**40.74/59.69**	**31.23/50.67**	**20.76/36.54**	**37.50/55.83**	**26.89/45.29**

Table 4: Verification experiment that our method alleviates the challenge of "partial robustness is distributed on the classifier".

Models	Datasets	Defense	Clean	FNA		SMA		IFGSM	
				8/255-16	10/255-16	8/255-16	10/255-16	8/255-16	10/255-16
ResNet50	Market to Duke	None	15.08/27.65	0.15/0.13	0.14/0.13	0.35/0.36	0.28/0.22	0.29/0.36	0.26/0.36
		Metric AT	16.51/29.35	4.47/10.89	2.96/9.07	10.77/22.52	9.49/21.14	5.78/13.14	4.53/12.07
		Ours	**19.07/34.69**	**6.17/12.88**	**4.65/10.23**	**13.02/24.60**	**11.71/22.80**	**7.92/15.35**	**6.46/13.33**
	Duke to Market	None	25.18/53.47	0.31/0.42	0.28/0.27	0.76/0.98	0.67/0.80	0.70/0.98	0.62/0.80
		Metric AT	19.54/45.25	6.27/17.67	4.50/12.80	13.32/30.43	11.57/26.51	7.50/20.61	5.89/16.00
		Ours	**21.87/50.12**	**7.87/21.91**	**6.20/18.05**	**14.16/34.32**	**12.73/30.46**	**9.67/25.83**	**8.09/21.85**

Table 5: Cross datasets evaluation to verify the generalization ability of our methods on ResNet50.

Ideally, the feature encoder E and the feature discriminator D should finally achieve the Nash equilibrium. At this point, the discriminator perceives all encoded features as either clean or adversarial features with probability of 0.5, meaning it cannot distinguish the origin of features at all. And the encoder extracts adversarial-invariant features shared between adversarial samples and clean samples. The above optimization process can be summarized as a min-max optimization framework:

$$\min_E \max_D \mathcal{L}(E, D) = \mathbb{E}_x[\log D(E(x))] + \mathbb{E}_{x^{\text{adv}}}\left[\log\left(1 - D\left(E\left(x^{\text{adv}}\right)\right)\right)\right], \tag{8}$$

where D and E are alternately optimized until D cannot distinguish features (i.e., Nash equilibrium).

Self-meta learning. To learn the generalized invariant features shared between seen and unseen IDs, we introduce self-meta learning. Specifically, for a batch of clean samples $\tilde{x}$ from the target distribution $\mathcal{D}$ and their adversarial samples $\tilde{x}^{\text{adv}}$, we divide them into a simulated training distribution $\mathcal{D}_{\text{meta-train}}$ for meta-training process and a simulated testing distribution $\mathcal{D}_{\text{meta-test}}$ for meta-testing process. The model first computes the loss function on $\mathcal{D}_{\text{meta-train}}$, and then updates its parameters via one-step gradient descent to obtain a temporary model G_{temp}:

$$\theta_G^{\text{temp}} = \theta_G - \alpha \nabla_{\theta_G} \mathcal{L}_{\text{meta-train}}\left(\mathcal{D}_{\text{meta-train}}\right), \tag{9}$$

where α is the temporary learning rate, and $\mathcal{L}_{\text{meta-train}}$ is defined as:

$$\mathcal{L}_{\text{meta-train}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{meta-train}}}\left[\ell(G, x, y) + \ell(G, x^{\text{adv}}, y^{\text{adv}})\right]. \tag{10}$$

where x^{adv} and y^{adv} is the corresponding adversarial sample and label of x and y. And $\ell = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{tri}} + \mathcal{L}_E$ is the total loss function. Subsequently, we use the same loss function as the meta-training phase to evaluate G_{temp} on the meta-testing data and obtain the meta-test loss:

$$\mathcal{L}_{\text{meta-test}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{meta-test}}}\left[\ell\left(G_{\text{temp}}, x, y\right) + \ell(G_{\text{temp}}, x^{\text{adv}}, y^{\text{adv}})\right]. \tag{11}$$

Finally, by integrating the losses from the meta-training and meta-testing phases, we obtain the overall loss for the self-meta learning:

$$\mathcal{L}_{\text{self-meta learning}} = \mathcal{L}_{\text{meta-train}} + \mathcal{L}_{\text{meta-test}}. \tag{12}$$

Notice that gradient descent is applied directly to θ_G based on $\mathcal{L}_{\text{self-meta learning}}$, instead of sequentially performing gradient descent on the meta-training data followed by the meta-testing data. The optimization can be formalized as:

$$\begin{gathered}\theta_G \leftarrow \theta_G - \beta \nabla_{\theta_G} \mathcal{L}_{\text{self-meta learning}}(\theta_G) \\ \Updownarrow \\ \min_{\theta_G}\left[\mathcal{L}_{\text{meta-train}}(\theta_G) + \mathcal{L}_{\text{meta-test}}\left(\theta_G - \alpha \nabla_{\theta_G} \mathcal{L}_{\text{meta-train}}(\theta_G)\right)\right],\end{gathered} \tag{13}$$

where β is the final learning rate. Intuitively, the meta-learning process evaluates whether the update direction of the meta-training phase has adaptive capability for the meta-testing data. If the adaptive capability is poor, the loss function value of the intermediate single-step updated model parameter θ_{temp}^G on the meta-testing data will be large, and this larger meta-testing loss corrects the update direction in return for better adaptation ability.

Additionally, the inherent second-order optimization scheme of meta-learning implicitly regularizes the model's first-order gradients, resulting in better generalization. To illustrate, we use the gradient chain rule to expand the gradient of the loss function in the meta-testing phase as an explicit function of the initial parameters θ_G:

$$\begin{aligned}&\nabla_{\theta_G} \mathcal{L}_{\text{meta-test}}\left(\theta_G^{\text{temp}}\right) \\ &= \nabla_{\theta_G^{\text{temp}}} \mathcal{L}_{\text{meta-test}}\left(\theta_G^{\text{temp}}\right) \cdot \nabla_{\theta_G}\left(\theta_G - \alpha \nabla_{\theta_G} \mathcal{L}_{\text{meta-train}}\left(\theta_G\right)\right) \\ &= \underbrace{\nabla_{\theta_G^{\text{temp}}} \mathcal{L}_{\text{meta-test}}}_{\text{first-order gradient}} - \alpha \underbrace{\nabla^2_{\theta_G} \mathcal{L}_{\text{meta-train}} \cdot \nabla_{\theta_G^{\text{temp}}} \mathcal{L}_{\text{meta-test}}}_{\text{hessian matrix}},\end{aligned} \tag{14}$$

where the second-order term $\nabla^2_{\theta_G} \mathcal{L}_{\text{meta-train}}$ implicitly regularizes the changing rate of the first-order gradient $\nabla_{\theta_G} \mathcal{L}_{\text{meta-train}}$ as:

$$\begin{aligned}&\text{Parameter update direction} \\ &\propto \nabla_{\theta_G} \mathcal{L}_{\text{meta-train}} - \alpha \nabla^2_{\theta_G} \mathcal{L}_{\text{meta-train}} \cdot \nabla_{\theta_G^{\text{temp}}} \mathcal{L}_{\text{meta-test}}.\end{aligned} \tag{15}$$

Here, the second-order term penalizes the curvature of the loss function (i.e., the rapid changes in the gradient), forcing the optimization path to be smoother and avoiding sharp local extrema, thus bringing better generalization.

Figure 5: Visualization of partial augmented data. The images with green borders are the original samples, while those with orange borders are the generated augmented samples.

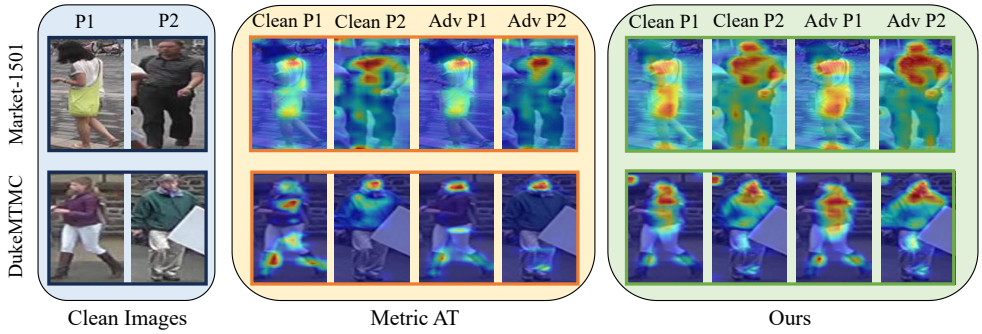


Figure 6: Heatmap based Grad-Cam. P1 means the 1_{st} person example.

Overall Model

By integrating the data balancing and the bi-adversarial self-meta defense, we obtain the final model. The intuitive framework can be referenced in Figure 4. The pseudocode of our method is provided in the supplementary materials.

Experiments

Basic Setup

Datasets and backbones. Following DAS (Wei et al. 2024), two common datasets, DukeMTMC and Market-1501, are used for evaluation. For the backbone networks, to ensure consistency, we also select ResNet50 (He et al. 2016) and APNet (Chen et al. 2021). Due to page limitations, please refer to the supplementary material for all evaluations of APNet. Following DAS, our adversarial training is conducted with FNA attack (Bouniot, Audigier, and Loesch 2020).

Evaluation metrics. Following the common setup of the person ReID (Zhou et al. 2023; Wei et al. 2024), we report the mean average precision (mAP) and Rank-1 accuracy of cumulative match characteristic (CMC). FNA, SMA (Bouniot, Audigier, and Loesch 2020) and Metric IFGSM (Bai et al. 2020) attacks are used for robustness evaluation. Table values follow the "mAP/Rank-1" format.

Compared defense. Due to the fundamental differences between classification tasks and retrieval tasks, most defense methods suitable for classification tasks do not apply to retrieval tasks. Thus, following DAS (Wei et al. 2024), we compare our method with the following defenses: BitSqueezing (Xu, Evans, and Qi 2018), MedianSmoothing2D (Xu, Evans, and Qi 2018), AMD (Bai et al. 2020), Adv_train (Bouniot, Audigier, and Loesch 2020), and DAS.

Training and evaluation details. Under our experimental setup, we run each experiment three times and report the best-performing result. For details regarding the software and hardware environments, hyperparameters of each module, attack settings, and training configurations, please refer to the supplementary materials.

Main Results

We report the main white-box robustness comparison on ResNet50 in Table 2. The results used for comparison are referenced from DAS (Wei et al. 2024). Our method achieves the current optimal adversarial robustness on the two datasets. To demonstrate its transferability, we also evaluate the black-box robustness in the supplementary material, which proves its transferable robustness.

Ablation Analysis

Firstly, we briefly analyze the effectiveness of each module based on ResNet50. Specifically, when none of our proposed modules is used, the method degenerates to FNA-based metric adversarial training baseline which is denoted as Metric AT. We first analyze the performance of each module when

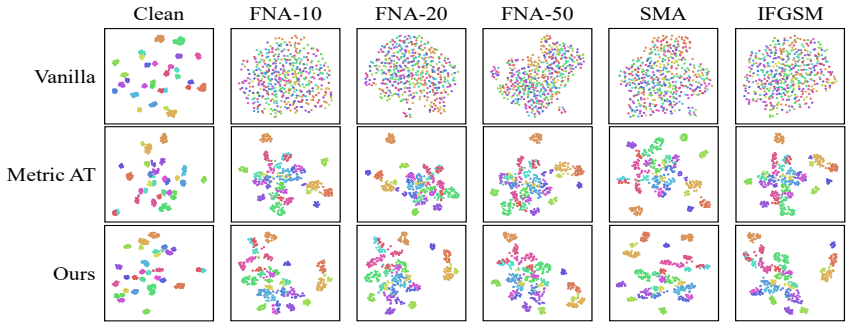


Figure 7: Feature distribution visualization based on UMAP.

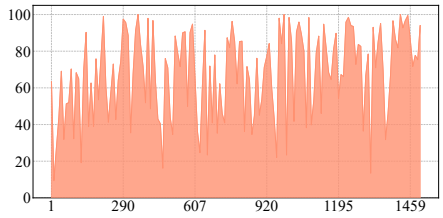


Figure 8: The mAP for each ID (avg=67.59, std.=22.74).

functioning individually, and then, we combine all modules to analyze the final model. The ablation study results are summarized in Table 3. Each module contributes a significant incremental benefit, and the optimal results are obtained when all modules are integrated.

Secondly, we follow the similar idea as in the introduction to verify that our method mitigates the challenges of model bias and robustness generalization. For the model bias challenge, the augmented data are shown in Figure 5, where IDs with fewer intra-class samples are filled to an average sample number level, and IDs lacking diversity are filled with isomeric pseudo data. We also statistically analyze the accuracy of our method across each IDs on Market-1501, as shown in Figure 8. Compared to Metric AT, our method exhibits lower inter-ID accuracy variance. For the generalization challenge, our method shows better robustness than Madry's AT (Madry et al. 2018) denoted as AT_PGD which uses PGD-based classification attack for adversarial training (shown in Table 4). This proves that our method alleviates the impact of absent classifiers. The evaluation of different attack methods in Table 2 preliminarily validates the generalization ability against unseen attacks. We also reference evaluation protocols for domain adaptation (Mekhazni et al. 2020; Feng et al. 2021) and domain generalization (Lin, Li, and Kot 2020; Liu, Ye, and Du 2024) to assess the generalization of our method on unseen domains and IDs. Specifically, we evaluate the robustness of the model trained on dataset A (e.g., Market) when tested on dataset B (e.g., Duke). Results shown in Table 5 demonstrate significant superiority of our defense on unseen domains, as the goals of adversarially-enhanced learning and self-meta learning modules explicitly induce the model to learn the dual-invariant representation.

Qualitative Analysis

We further validate the effectiveness of the proposed method through qualitative analyses in this section.

Heatmap. The heatmap reflects high attention areas of the model to the samples. A model with strong discriminative ability should focus on key regions within the samples, such as the contours and body of a person. We visualize the heatmaps of both the Metric AT model and our proposed model using Grad-CAM (Selvaraju et al. 2017), as shown in Figure 6. Visually, our method restores the model's attention areas on adversarial samples, focusing on regions with higher discriminative information. The heatmap of our method is more interpretable than that of the Metric AT model since it is more in line with human contours. This provides interpretability of adversarial robustness from the perspective of the model's preferences.

Feature distribution visualization. Intuitively, a model with strong discriminative ability should encode features with smaller intra-class distances and larger inter-class distances. We use UMAP (McInnes, Healy, and Melville 2018) to visualize the feature space distributions of the vanilla trained model, the Metric AT model, and our proposed defense model in Figure 7. Visually, the encoding features of the vanilla model for adversarial samples are scattered in the feature space, indicating poor discriminability. The Metric AT model maintains a discernible feature distribution for adversarial samples and our method demonstrates tighter intra-ID clustering and greater inter-ID distance compared to Metric AT. This provides the interpretability on adversarial robustness from the perspective of feature distribution.

Conclusion

In this paper, we analyze the two challenges ignored by current adversarial defenses in person ReID, i.e., model bias and composite generalization requirements. We then propose a novel debiased dual-invariant defense to address these challenges. we address model bias through data balancing and address composite generalization requirements through a bi-adversarial self-meta defense framework. Our defense achieves the optimal person ReID adversarial robustness, and the extended experiments and ablation analysis demonstrate its effectiveness, transferability, and interpretability. We hope our work can draw attention to trustworthy and robust person ReID models within the community.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China 62572150 and 62372137, in part by the Shenzhen Science and Technology Program KJZD20230923114806014 and JCYJ20230807094411024, in part by the Guangdong Basic and Applied Basic Research Foundation 2024A1515012299, and in part by the Major Key Project of PCL PCL2024A05.

References

Bai, S.; Li, Y.; Zhou, Y.; Li, Q.; and Torr, P. H. 2020. Adversarial metric attack and defense for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6): 2119–2126.

Bian, Y.; Liu, M.; Wang, X.; Ma, Y.; and Wang, Y. 2025. Learning to learn transferable generative attack for person re-identification. *IEEE Transactions on Image Processing*.

Bouniot, Q.; Audigier, R.; and Loesch, A. 2020. Vulnerability of person re-identification models to metric adversarial attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 794–795.

Chen, G.; Gu, T.; Lu, J.; Bao, J.-A.; and Zhou, J. 2021. Person Re-Identification via Attention Pyramid. *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 30: 7663–7676.

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, 2206–2216. PMLR.

Feng, H.; Chen, M.; Hu, J.; Shen, D.; Liu, H.; and Cai, D. 2021. Complementary pseudo labels for unsupervised domain adaptation on person re-identification. *IEEE Transactions on Image Processing*, 30: 2898–2907.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the Design Space of Diffusion-Based Generative Models. In *Advances in Neural Information Processing Systems*, volume 35, 26565–26577.

Lin, S.; Li, C.-T.; and Kot, A. C. 2020. Multi-domain adversarial feature generalization for person re-identification. *IEEE Transactions on Image Processing*, 30: 1596–1607.

Liu, F.; Ye, M.; and Du, B. 2024. Domain generalized federated learning for person re-identification. *Computer Vision and Image Understanding*, 241: 103969.

Liu, X.; Kuang, H.; Lin, X.; Wu, Y.; and Ji, R. 2023. CAT: Collaborative adversarial training. *arXiv preprint arXiv:2303.14922*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mekhazni, D.; Bhuiyan, A.; Ekladious, G.; and Granger, E. 2020. Unsupervised domain adaptation in the dissimilarity space for person re-identification. In *European Conference on Computer Vision*, 159–174. Springer.

Raghunathan, A.; Xie, S. M.; Yang, F.; Duchi, J. C.; and Liang, P. 2019. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2018. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*.

Wang, Z.; Zheng, S.; Song, M.; Wang, Q.; Rahimpour, A.; and Qi, H. 2019. advpattern: Physical-world attacks on deep person re-identification via adversarially transformable patterns. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8341–8350.

Wei, J.; Pan, C.; He, S.; Wang, G.; Yang, Y.; and Shen, H. T. 2024. Towards robust person re-identification by adversarial training with dynamic attack strategy. *IEEE Transactions on Multimedia*, 26: 10367–10380.

Wei, X.-S.; Song, Y.-Z.; Mac Aodha, O.; Wu, J.; Peng, Y.; Tang, J.; Yang, J.; and Belongie, S. 2021. Fine-grained image analysis with deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12): 8927–8948.

Xu, W.; Evans, D.; and Qi, Y. 2018. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *Proceedings 2018 Network and Distributed System Security Symposium*. Internet Society.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 2872–2893.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, 7472–7482. PMLR.

Zhang, J.; Xu, X.; Han, B.; Niu, G.; Cui, L.; Sugiyama, M.; and Kankanhalli, M. 2020. Attacks which do not kill training make adversarial learning stronger. In *International conference on machine learning*, 11278–11287. PMLR.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. *arXiv preprint arXiv:1609.01775*.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable Person Re-identification: A Benchmark. In *IEEE International Conference on Computer Vision (ICCV)*.

Zheng, Z.; Zheng, L.; Yang, Y.; and Wu, F. 2023. U-turn: Crafting adversarial queries with opposite-direction features. *International Journal of Computer Vision*, 131(4): 835–854.

Zhou, Y.; Huang, F.; Chen, W.; Pu, S.; and Zhang, L. 2023. Stochastic gradient perturbation: An implicit regularizer for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10): 5894–5907.