Person Re-identification Method Based on Color Attack and Joint Defence

Yunpeng Gong   Liqing Huang*   Lifei Chen†
College of Computer and Cyber Security, Fujian Normal University, P. R. China
fmonkey625@gmail.com   {lqhuang,clfei}@fjnu.edu.cn

Abstract

The main challenges of ReID is the intra-class variations caused by color deviation under different camera conditions. Simultaneously, we find that most of the existing adversarial metric attacks are realized by interfering with the color characteristics of the sample. Based on this observation, we first propose a local transformation attack (LTA) based on color variation. It uses more obvious color variation to randomly disturb the color of the retrieved image, rather than adding random noise. Experiments show that the performance of the proposed LTA method is better than the advanced attack methods. Furthermore, considering that the contour feature is the main factor of the robustness of adversarial training, and the color feature will directly affect the success rate of attack. Therefore, we further propose joint adversarial defense (JAD) method, which includes proactive defense and passive defense. Proactive defense fuse multi-modality images to enhance the contour feature and color feature, and considers local homomorphic transformation to solve the over-fitting problem. Passive defense exploits the invariance of contour feature during image scaling to mitigate the adversarial disturbance on contour feature. Finally, a series of experimental results show that the proposed joint adversarial defense method is more competitive than a state-of-the-art method.

1. Introduction

Person re-identification (ReID) is matching the same person across diferent cameras and scenes [1–4]. This technology have been widely applied to video surveillance [5–7], image retrieval [8, 9], criminal investigation [8], target tracking [10] and others. ReID has been a challenging and hot problem since illumination, complex environment, occlusion, image blur and other factors. In recent years, many ReID works [5–10] used deep-learning module, and have made great progress. However, Szegedy et al. [11] found the deep-learning models are susceptible to attacks from adversarial samples, which will cause the network to completely change its prediction results. The works of [12–16] have proved that the ReID systems based on deep-learning have the same vulnerability. And these adversarial samples have only added a slight disturbance, which is hidden enough for the human visual system. It is very important to study the security of ReID systems because the insecurity may cause severe losses, eg., criminals may use adversarial disturbance to cheat the monitoring systems [14, 16].

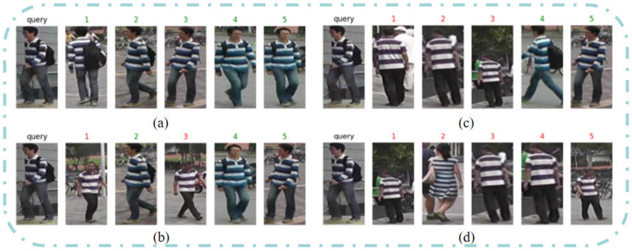


Figure 1. (a) shows the retrieval results of clean example. (b) corresponds to Meric-IFGSM attack [12], (c) corresponds to the SMA attack [13], (d) corresponds to the proposed LTA attack. The numbers on the images indicate the rank of similarity in the retrieval results, the red and green number denote the wrong and correct results, respectively.

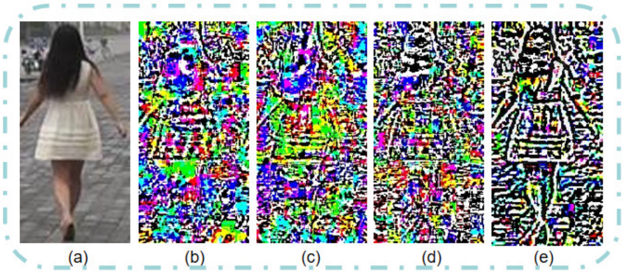


Figure 2. (a) shows clean example and (b) (c) (d) shows the adversarial noise generated after attacking three different models using SMA [13], respectively. (b) corresponds to the normally trained model, (c) corresponds to the model which using [17] to train with better robustness to color variations, (d) corresponds to our proactive defense model, (e) corresponds to our joint adversarial defense.

*Equal contribution
†Corresponding author

The adversarial metric attack usually requires additional push or pull guidance to distort the distance between the attacked image and other images with the same identity or class [12–14, 16, 18–23], so as to achieve the purpose of deception models. Therefore, most of the existing studies [12, 13, 18, 23] of adversarial metric attack and defense revolve around metric relationships. Generally, the optimization function of the adversarial noise is designed for pulling the distance between negative pairs and pushing the distance between positive pairs.

The intra-class variations caused by color deviations such as lighting, chromatic aberration, etc., in the various camera conditions is one of the main challenge for ReID [24–26], due to the training set solely encompasses a limited portion of the intra-class variations of the color domain, the model is easy to overfit. In the existing adversarial metric attacks [12–14], it was observed that the attacks naturally perturb the color feature of the samples, which consistent with the color feature is important clue for image retrieval. The effect of the adversarial attack on the color feature is visualized in Figure 1. When retrieving clean samples, the model was able to identify the retrieved pedestrians with blue-gray striped top, grayish trousers, and backpack, even though they looked a little off from different cameras. While under adversarial perturbation, as shown in (b), (c), (d), the misjudgment of the model in color becomes more and more serious, which includes the colors of pedestrian top, trousers and backpack. Two classic metric attacks include metric-IFGSM [12] and SMA [13] attack. The metric-IFGSM [12] attack was realized by maximizing the metric distance between the retreved image and the other images with the same identity, and used reference images. The SMA [13] attack added random noise to the retrieved image, and maximizing the metric distance from the clean image. The SMA not require reference images and thus is more with realistic scenario. Compared with metric-IFGSM, SMA only used the retrived image to generate disturbance, so it performs better in making the model misjudge color. Therefore, we further propose loacal transformation attack (LTA), which does not add random noise, but use local gray transformation with more obvious color variation to randomly disturb the color of the retrieved image, so as to learn robust adversarial noise against the color variation, and to strengthen the attack on color feature. Finally, experiments verify that the performance of the proposed LTA method is better than the advanced methods.

After exploring the vulnerability of ReID attack, we begin to research the effectiveness of defense methods. Adversarial training is currently the main adversarial metric defense method [12, 13, 18, 23]. Generally speaking, in adversarial training, a defense model trained by adversarial examples of an attack cannot defend against multiple attacks at the same time [27], and extreme overfitting during training leads to obvious reduction in model generalization capacity [28].

Considering that being better at capturing shape or contour features is the main factor for the robustness of adversarial training [29], and color features have a direct impact on the success rate of attacks. So there speculate that color features and contour feature are inherently important targets for attacker. To this end, we propose a corresponding joint adversarial defense approach. Firstly, we consider increasing the robustness of the model to color variations as a proactive defense. We speculate that when the robustness of the model to color variations is increased, the adversary will change the attack direction and strengthen the attack on the contour feature. It can be seen from Figure 2(c) that the contour feature have been more seriously damaged. In addition, we also fuse sketch images during the model training process to strengthen the learning of contour feature so as to enhance the defense against the two attack modes (color and contour). From Figure 2(d), it can be seen that the adversarial noise is significantly weakened on our proactive defense model. And then, we further propose a passive defense strategy, which utilizing the invariance of contour features in the circuitous scaling to mitigate attack on contour feature. This strategically complementary to the proactive defense. From Figure 2(e), it can be seen that after implementing the joint adversarial defense proposed in this paper, the adversarial noise becomes very sparse, and the contour feature are also well protected. The proposed joint defense model is a lightweight method without any additional parameter learning. It can be combined with various ReID models without changing the learning strategy. Therefore, the main contributions of this paper are summarized as follows:

- We propose a new attack method – local transformation attack (LTA) for the first time, by using more obvious color variation to randomly disturb the color of the retrieved image and without reference image.
- We propose a joint adversarial defense model based on feature-invariant is to against adversarial metric attacks, which does not rely on adversarial training. The proposed method improves the robustness of the model, and performs well in cross-domain tests.
- Finally, the comparative experimental results with the state-of-the-art algorithms further verified the effectiveness and the advanced nature of the proposed method.

2. Related Work

In this section, the previous work on adversarial attacks and defenses of the metric learning is described.

2.1. Adversarial Attacks

Adversarial attacks can be categorized into white-box [12, 30] and black-box [14, 31] attacks. The black-box attack means that the attacker does not know the structure and parameters of the target network, and the adversaries can only resort to the query access to generate adversarial samples. White-box attack assumes that the attacker has prior knowledge of the target networks, including the structure and parameters of model, which means that the adversarial examples are generated with and tested on the targe network. For the same attack, the success rate of white-box attack is higher than black-box attacks.

There are some metric attack methods proposed in ReID. Metric-FGSM [12] extended some metric attacks by classification attacks, including fast gradient sign method (FGSM) [32], iterative FGSM (IFGSM) and momentum IFGSM (MIFGSM) [33]. Among the three attack methods, IFGSM delivers the strongest white-box attacks [12]. Opposite-direction feature attack (ODFA) [15] exploits feature-level adversarial gradients to generate adversarial examples to pull the feature in the opposite direction with an artificial guide. Self metric attack (SMA) [13] uses the image with added noise as the reference image and obtains the adversarial examples by attacking the feature distance between the original image and the reference image. This process does not require any additional images, it is more in line with the actual situation that the attacker usually lacks data. Furthest-negative attack (FNA) [13] combine hard sample mining [34, 35] and triple loss to obtain pushing guides and pulling guides to move image feature to head towards the least similar cluster of features while moving away from the other similar features. Deep mis-ranking (DMR) [14] proposed a learning-to-mis-rank formulation to perturb the ranking of the system output, which used a multi-stage network architecture that pyramids the features of different levels to extract general and transferable features for the adversarial perturbations. The success attack rate of black-box attacks is almost as high as that of white-box attack. At the same time, it also showed that when applied to classification attacks, it has a higher success attack rate than DeepFool [36], NewtonFool [37], and CW [38], and it has successfully broken through many classical ReID models [25, 39–47].

2.2. Adversarial Defenses

Recently, a number of effective defense methods have been employed to against adversarial classification attacks [48–54], such as denoising methods, randomization-based schemes, adversarial training and others. The defense methods based on denoising, such as Guo et al. [48] used more diversified non-differentiable image transformation operations, which includes depth reduction, total variance minimization and image quilting. The goal is to increase the difficulty of network gradient prediction, and then achieve the purpose of defense. Noting that most of the training images are in JPG format, Dziugaite [49] used JPG image compression method to reduce the impact of adversarial disturbance. In terms of randomization, RRP (random resizing and padding) [50] mitigates adversarial effects by combining random resizing and random padding based on adversarial training. [51–54] showed that adversarial training is a robust way to defend against adversarial attacks, which includes offline adversarial training and online adversarial training.

The metric defense schemes employ by [12, 13, 18, 23] correspond to offline adversarial training and online adversarial training respectively. The defense method uses in [12] is offline adversarial training, which is based on a generation of an adversarial version of the training set obtained with a frozen version of the trained model. As a frozen model is used to generate attacks, this method is referred to as offline adversarial training. The defense method uses in [13, 18, 23] is online adversarial training, which generates adversarial examples online while the defended model evolves by triplet loss. However, adversarial training is prone to overfitting [13, 28] because dependent on the training data results in reducing the generalization capacity of the model. Enhancing the robustness towards adversarial examples and maintaining the generalization capacity of the model is the important issue of adversarial defense.

3. Proposed Methods

In this section, we propose the local transformation attack (LTA) based on color features. In order to push the feature of the reference image away from the original image, there constructs a reference image with local difference from the original image in each basic iteration based on LGT [17]. As for the proposed attack method (LTA), we further propose a joint defense method. In proactive defense, we combine the three modal images of visible (RGB), grayscale and sketch for random channel fusion. In passive defense, it realizes by circuitous scaling of image. The specific attack and joint defende method framework is showed in Figure 3.

3.1. Proposed Local Transformation Attack

In order to attack the color feature, we propose the local transformation attack (LTA), which adopts local grayscale transformation (LGT) [17] constructing the local color deviation of the input. And then, it randomly selects a rectangular area in the image and replaces it with the pixels of the same rectangular area in the corresponding grayscale image. As showed in Figure 3, the LGT makes the constructed reference image have appropriate local differences from the original image.

The initialization of the proposed LTA method is defined as:

$$x_{adv}^{(0)} = x \tag{1}$$

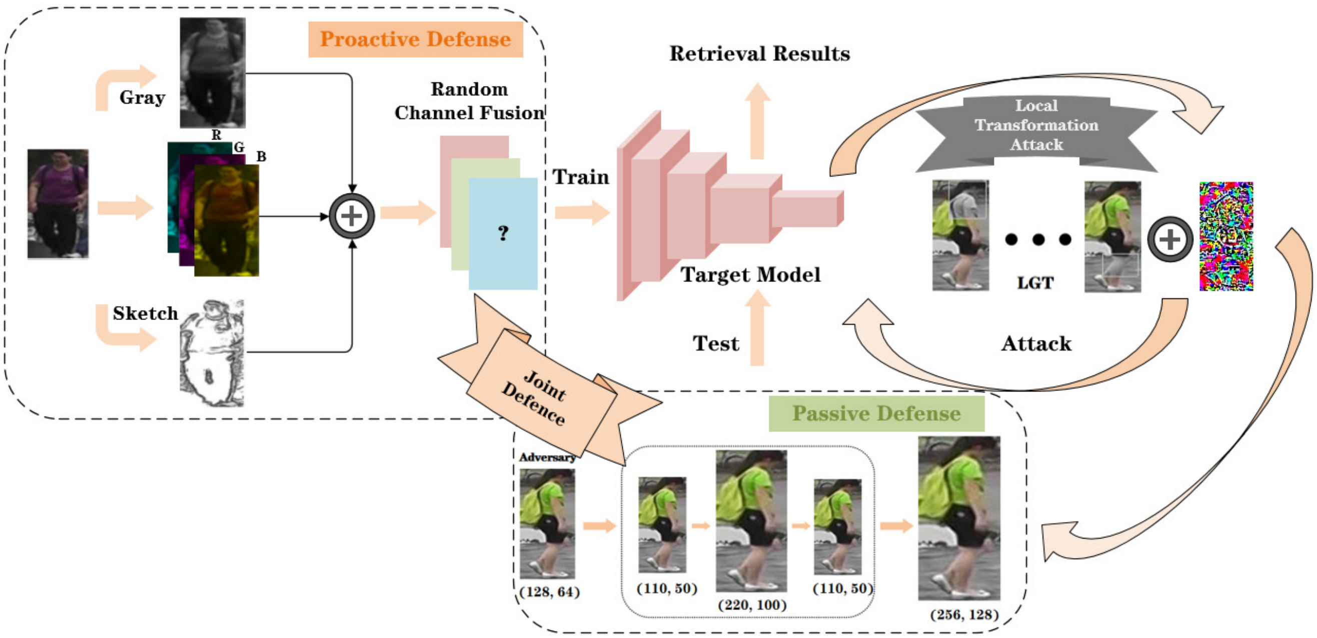


Figure 3. Framework diagram of our attack and joint adversarial defence. In our Local Transformation Attack (LTA), It pushes the feature of the reference image away from the original image by constructing a reference image with local difference from the original image in each basic iteration based on LGT [17]. In proactive defense, we combines the three modal images of visible (RGB), grayscale and sketch for random channel fusion. In our passive defense, it realizes by circuitous scaling of image.

where x denote the attacked image. There using $x_{adv}^{(n)}$ denotes the adversarial example at the n-th iteration. $\hat{x}^{(n)}$ is the reference image with local variability constructed by LGT [17] at the n-th basic iteration. So the proposed LTA is defined as the following iterative optimization:

$$\hat{x}^{(n)} = LGT(x) \tag{2}$$

$$x_{adv}^{(n+1)} = \Psi_x^{\varepsilon}(x_{adv}^{(n)} + \alpha \cdot sign(grad^{(n+1)})) \tag{3}$$

where ϵ is the adversarial bound and α is the iteration step size, Ψ_x^{ε} is the clip function, which ensures that $\|x_{adv}^{(n+1)} - x\|_\infty < \epsilon$ and that adversarial noise inconspicuousness. And the $grad^{(n)}$ is the accumulated gradient at the n-th iteration:

$$grad^{(n+1)} = \theta \cdot grad^{(n)} + \frac{\Delta_{LTA}^{(n)}}{\|\Delta_{LTA}^{(n)}\|_1} \tag{4}$$

where θ is the decay factor of the momentum term, in our experiments θ is set to 1. And $\Delta_{LTA}^{(n)}$ is calculated as follows:

$$\Delta_{LTA}^{(n)} = \frac{\partial D(f_{adv}^{(n)}, \hat{f}^{(n)})}{\partial f_{adv}^{(n)}} \tag{5}$$

$$D(f_{adv}^{(n)}, \hat{f}^{(n)}) = \|f_{adv}^{(n)} - \hat{f}^{(n)}\|_2^2 \tag{6}$$

where $f_{adv}^{(n)}$ denotes the feature of the adversarial example. Specifically, each iteration optimizes adversarial noise by attacking the feature distance between the adversarial image generated from the result of previous iteration and the new reference image.

3.2. Proposed Joint Adversarial Defence Method

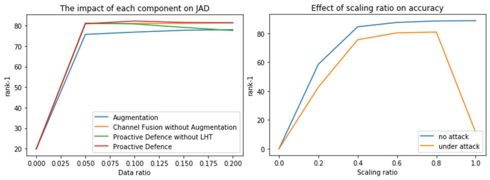
In order to overcome the attack based on color features, we further propose the joint adversarial defense method (JAD). The proposed method includes the proactive and the passive defense. The proactive defense consists of channel fusion (CF) and local homogeneous transformation (LHT), and the passive defense consists of circuitous scaling (CS).

3.2.1 Proposed Proactive Defence

The proactive defense consists of channel fusion (CF) and local homogeneous transformation (LHT).

Channel fusion (CF). Visible images, grayscale images, and sketch images are homogeneous, which contain the same structural information. The results in [17, 55] showed that using the homogeneous grayscale images to learn structural information in training is effective in increasing the robustness to color variations.

We add grayscale information and sketch information by channel fusing. The operation (Grayscale(3)) in Pytorch is adopted to get the grayscale image for each visible image, and sketch images can be obtained by inverting the grayscale image and then Gaussian blurring it, finally blending it with the grayscale image. As shown in Figure 3, the RGB images are randomly converted with a certain



(a) (b)

Figure 4. Hyper-parameter sensitivity analysis: (a) the contribution of each component of the joint defense to the defense and the impact of different ratios on defense performance; (b) the effect of different image scaling ratios on defense performance in passive defense.

probability into 3-channel grayscale images or sketch images in pre-processing stage, and then randomly merge the channels of the grayscale image and the sketch image with the channels of the RGB image to create a new homogeneous modal image.

In the process of CF, 1 or 2 channels are randomly selected from the R, G, and B channels of the visible image. After the visible image channel and the number of channels n is determined, the grayscale or sketch image channel is randomly selected to reconstruct a new 3-channel image. In fact, a maximum of 60 homogeneous variations can be generated by combining the 5 image channels types of R, G, B, grayscale, and sketch in random order.

In addition, from Figure $4a$, we can see that the augmentation based on image transformation will further extend the diversity of modes, such as Posterize, Equalize, Solarize, Contrast, Inversion and so on. However, the multi-modal inputs will lead to overfitting of the model, which affects the generalization capacity of the model as showed in Figure $5a$.

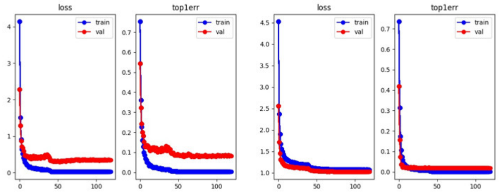
Local homogeneous transformation (LHT). To solve this problem, we propose a strategy method based on local homogeneous transformation (LHT), which extends local grayscale transformation (LGT) [17]. Using LHT to guide the model to fit the diversity of variation gradually from local variations. The difference of LGT is that the proposed LHT replaces randomly selected regions with homogeneous images. As showed in Figure $5b$, it positively helps reduce the overfitting in training. Unless otherwise specified, the diversity data learning in subsequent experiments combine with LHT by default.

The LHT for each visible image x^v can be achieved by the following equations:

$$x^h = T(x^v), \tag{7}$$

$$rect = RandPosition(x^v), \tag{8}$$

$$x^{LHT} = LT(x^v, x^h, rect) \tag{9}$$



(a)Training normally (b)Training with LHT

Figure 5. The comparison of training curve without LHT and with LHT.

and

$$(x^{LHT}|y) = (x^v|y) \tag{10}$$

where x^h is the homogeneous images, and $T(\bullet)$ is the homogeneous transformation funtion; $RandPosition(\bullet)$ is used to generate a random rectangle in the image, and the function of $LT(\bullet)$ is to give the pixels in the rectangle corresponding to the x^h image to the x^v image; x^{LHT} is the sample after local homogeneous transformation, and y is the label of the transformed image.

From Figure $4a$, we can see that the best defensive performance is achieved when the components are in the ratio of 5% to 15%. Therefore, the probability of the image using augmentation transformation is set to 5%, and the probability of converting to a grayscale image is set to 5%. In addition, the probability of using CF transformation is set to 5%, and the probability of using LHT is set to 10%.

3.2.2 Proposed Passive Defence

Since color features and contour feature are two important targets in the attack, the increased robustness of the model to color variations will force the adversary to change the direction of the attack to some extent, more towards attacking contour feature. Therefore, we exploit the invariance of contour features during image scaling to mitigate the adversarial disturbance on contour feature.

The basic principle of image scaling is to calculate the pixel value of the target image according to the pixel value of the original image by certain rules, common image scaling algorithms such as linear interpolation [56]. In the scaling process, some pixels are discarded or some new pixels are added. [50] found that the adversarial noise structure can be effectively destroyed by one-time image scaling. Circuitous scaling (CS) consists of multiple image scaling to give full play to this advantage.

The passive defense is realizes by a series of image resizing. The scaling of an image does not bring more information about the image, so the quality of the image will inevitably be affected, which also has an impact on the retrieval accuracy. Therefore, it is important to find a suitable scaling ratio to trade-off the retrieval accuracy and the adversarial robustness. The effect of scaling ratio on accuracy can be seen in Figure $4b$. Taking the Market1501 [2] dataset

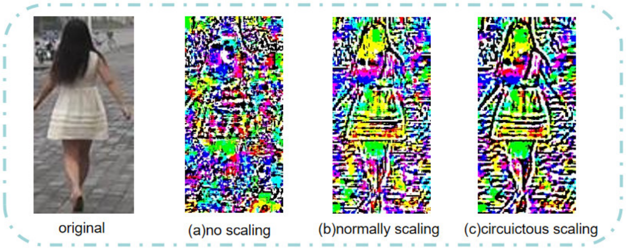


Figure 6. The effect of image scaling and CS on the adversarial noise, where (a) shows the adversarial noise from the original adversarial example, and (b) shows the adversarial noise after reszie the adversarial sample to [110, 50] (then restored to [256, 128]). (c) shows the adversarial noise after CS.

as an example, the original size of the dataset image is [128, 64], and the size is uniformly resized to [256, 128] when fed into the CNN. When using image resizing as passive defense, we observe that the network performance hardly drops and gets a satisfactory defense effect if we resize the image to [110, 50] (Approximately 0.8 times the original image) to corrupt the adversarial noise structure. Our passive defense consists of a series of resizing that reszie the image to [110, 50] then to [220, 100], then to [110, 50] again (finally uniformly to the [256, 128]), so it called circuitous scaling. The effect of image scaling [50] (only once resizing) and CS on the adversarial noise can be seen in Figure 6, and it can be observed that the adversarial noise at the contour feature is continuously weakened, and the outline of the pedestrian is more clearer.

4. Experiments

In this section, we evaluate our LTA by comparing with SMA [13] and then evaluate the robustness of our approach using cross-domain tests. Finally, we verify the effectiveness of our JAD under white-box attacks and black-box attack.

4.1. Attack Evaluation and Cross-Domain Tests

Datasets. Experiments are conducted on Market1501 [2] and DukeMTMC [45]. The Market1501 includes 1,501 pedestrians captured by six cameras (five HD cameras and one low-definition camera). The DukeMTMC is a large-scale multi-target, multi-camera tracking dataset, a HD video dataset recorded by 8 synchronous cameras, with more than 2,700 individual pedestrians. The above two datasets are widely used in ReID studies.

Evaluation criteria. Following existing works [2], Rank-k precision and mean Average Precision (mAP) are adapted as evaluation metrics. Rank-1 denotes the average accuracy of the first return result corresponding to each query image. mAP denotes the mean of average accuracy, the query results are sorted according to the similarity, the closer the correct result is to the top of the list, the higher the score.

Table 1. Evaluation on Market1501 [2] under a white-box attack on the query set. Where LTA* means that only one version with local differences is used as a reference image, and LTA generates image versions with different local differences in each basic iteration to conduct the metric attack.

Attack	Rank-1	Rank-5	Rank-10	mAP
No-attack	88.4%	95.5%	97.1%	72.1%
SMA [13]	15.7%	26.4%	32.7%	11.1%
LTA*(ours)	15.7%	26.2%	32.1%	11.0%
LTA(ours)	**13.3%**	**22.4%**	**28.1%**	**9.6%**

Implementation details. The proposed adversarial attack and defense algorithm is development based on on PyTorch framework. In our baseline, ResNet50 [57] and DenseNet [40] are used as the backbone network in experiments, and the pre-trained ImageNet parameters are adopted for network initialization. Specifically, the stride of the last convolutional block is set to 2. We adopt the stochastic gradient descent (SGD) optimizer for optimization, and the momentum parameter is set to 0.9. We set the initial learning rate as 0.1. The learning rate is decayed by 0.1 every 40 iteration, with a total of 60 training epochs and a batch size of 32 for normal training on both datasets, and 120 training epochs for our method as well as for adversarial training [12].

Attack evaluation. The hyper-parameters are unified for fair comparison, the adversarial boundary is set to 5 pixels, the iteration step size is set to 1, and the number of basic iterations is set to 15. Note that in contrast to an adversarial defense problem, lower precision indicates better attack performance. It can be seen from Table 1 that when only one version of the reference image is used, the success attack rate of LTA* is better than that of SMA. The comparison between LTA* and LTA shows that using diverse versions of reference images has a higher attack success rate than using only one version of reference images. The experimental results fully demonstrate that the attack against color features are more aggressive compared to the same type of SMA attack.

Cross-domain tests. It is pointed out by [12] that the higher accuracy of the model does not mean that it has better generalization capacity. The defense capabilities of different baselines under the same attack would have been greatly different, and the high accuracy model may even have worse defenses capabilities due to overfit. In response to the above potential problems, we suggest to use cross-domain tests and adversarial defense tests to verify the robustness of the model. Experiments show that the proposed method effectively enhances the generalization capacity of the model, and the Table 2 shows the cross-domain experiments of the proposed method between two datasets, Market-1501 [2] and DukeMTMC [45]. We use the state-of-the-art defense

Table 2. The performance of different models is evaluated on cross-domain dataset. M→D means that we train the model on Market1501 [2] and evaluate it on DukeMTMC [45].

Model	M→D		D→M	
	Rank-1	mAP	Rank-1	mAP
Baseline	36.1%	18.9%	45.7%	19.6%
GOAT [13]	23.6%	11.4%	47.3%	18.5%
JAD(w/o LHT)(ours)	36.9%	18.4%	47.4%	**19.5%**
JAD(ours)	**42.5%**	**21.5%**	**47.5%**	19.4%

Table 3. The performance of normally trained models (Baseline) and our JAD models on Market1501 and DukeMTMC.

Methods	Market1501 [2]		DukeMTMC [45]	
	Rank-1	mAP	Rank-1	mAP
Baseline	88.4%	72.2%	78.7%	62.3%
Baseline+RK [58]	90.2%	84.7%	83.3%	79.3%
JAD(ours)	88.7%	70.3%	77.2%	57.8%
JAD+RK(ours)	91.0%	85.0%	82.7%	77.0%

Table 4. The performance of normally trained models and our JAD models under white-box attack of the query.

Dataset	Model	Rank-1/mAP(%)		
		M-IFGSM [12]	SMA [13]	LTA
Market1501 [2]	Baseline	8.1/4.3	15.7/11.1	13.3/9.6
	Baseline+RK	13.2/13.0	17.6/20.0	14.2/16.1
	JAD(ours)	47.1/27.8	79.3/60.1	56.5/41.1
	JAD+RK(ours)	61.3/56.8	85.6/80.3	66.2/63.9
DukeMTMC [45]	Baseline	10.1/5.8	15.0/10.4	13.0/9.2
	Baseline+RK	16.8/16.3	18.8/19.8	15.2/16.3
	JAD(ours)	30.5/16.3	56.7/39.5	41.8/27.6
	JAD+RK(ours)	48.1/43.4	69.3/64.5	52.4/49.7

model GOAT [13] for comparison.

In the cross-domain tests of Market1501→DukeMTMC, it can be seen that JAD (without LHT) enhances the Rank-1 by 3.1 percentage points compared with the baseline, and further enhances by 4.8 percentage points after using LHT. In the cross-domain tests of DukeMTMC→Market1501, it can be seen that the proposed method (without LHT) enhances the Rank-1 by 3.4 percentage points compared with the baseline, and further enhances by 0.2 percentage points after using LHT. The above shows that the proposed method effectively enhances the generalization capacity of the model, and LHT further enhances the generalization capacity.

4.2. Experiments of JAD

This subsection verifies the effectiveness of the proposed method from white-box [12, 13] attack, black-box [14] attack and the other baselines [59,60], and shows the effect of each component of the proposed method in defense through the ablation experiment. Then, the state-of-the-art black-box attack DMR [14] is used to compare defence performance of our JAD and the state-of-the-art defense method GOAT [13]. Finally, we give a visual analysis of our defense.

Table 5. Comparison of baseline, channel fusion (CF), proactive defence (PD), and joint adversarial defense (JAD) in terms of defense accuracy under white-box attack on Market1501 [2].

Model (with RK)	Rank-1/mAP(%)			
	No-attack	M-IFGSM [12]	SMA [13]	LTA
Baseline	90.2/84.7	13.2/13.0	17.6/20.0	14.2/16.1
CF(ours)	90.8/85.3	18.3/16.7	18.0/19.6	17.1/18.5
PD(ours)	91.5/85.6	31.7/28.7	58.4/57.1	25.2/27.6
JAD(ours)	91.0/85.0	61.3/56.8	85.6/80.3	66.2/63.9

Table 6. Comparison of different resizing combinations in terms of defense accuracy, where P_1 means the scaling pattern that resize image to [110, 50] and P_2 means resize to [220, 100]. $P_1 \to P_2$ means resize image to [110, 50] then to [220, 100].

Dateset	Model	No-attack	LTA Attack
		Rank-1/mAP(%)	Rank-1/mAP(%)
Market1501 [2]	Baseline	90.2/84.7	14.2/16.1
	P_1	90.0/84.7	25.5/27.5
	P_2	90.3/84.9	15.9/17.9
	$P_1 \to P_2$	90.1/84.8	25.8/27.8
	$P_1 \to P_2 \to P_1$	89.8/84.3	**31.4/33.1**

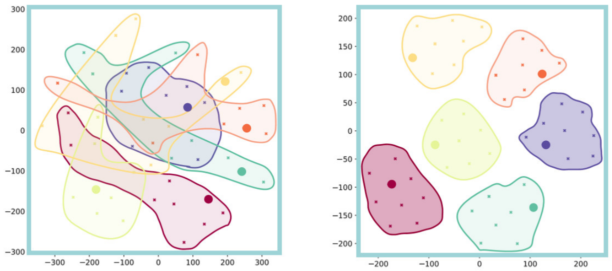
Table 7. Comparison of different defense methods in terms of defense accuracy under white-box attacks on Market1501.

Model	Rank-1/mAP(%)			
	No-attack	M-IFGSM [12]	SMA [13]	LTA
Baseline	90.2/84.7	13.2/13.0	17.6/20.0	14.2/16.1
AT [12]	86.4/76.9	46.8/41.3	48.3/47.4	49.1/48.4
AT +CS	86.3/76.1	60.6/53.8	64.9/61.1	62.7/58.5
JAD(ours)	90.6/84.3	**66.9/62.1**	**86.5/80.7**	**73.5/69.5**

Table 8. Comparison of different defense methods and other baselines in terms of defense accuracy under DMR black-box attack.

Model (w/o RK)	Rank-1/mAP(%)		Model (with RK)	Rank-1/mAP(%)	
	No-attack	DMR [14]		No-attack	DMR
Baseline	88.4/72.2	19.8/15.8	SB [59]	95.4/94.2	6.2/4.8
GOAT [13]	87.5/66.9	67.8/46.4	SB+JAD(ours)	95.1/94.0	93.3/91.2
GOAT+CS	88.0/68.3	72.8/50.7	FR [60]	96.8/95.3	24.8/25.9
JAD(ours)	88.7/70.3	81.1/60.7	FR+JAD(ours)	96.3/94.9	91.6/90.1

We tested our JAD with white-box attacks on Market1501 [2] and DukeMTMC [45], and the attacks include metric-IFGSM (M-IFGSM) [12], SMA [13] and the proposed LTA. To be consistent with recent works, we follow the new training/testing protocol to conduct our experiments by k-reciprocal re-ranking (RK) [58]. It can be seen from Table 4 that on the two datasets, our JAD has enhanced the Rank-1 in all white-box attacks by more than 40 percentage points after using re-ranking.



(a)Normally trained model

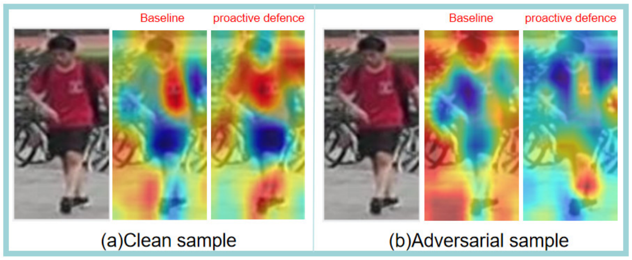
(b) Proactive defense model

Figure 7. t-SNE [61] visualization of six randomly selected images with different identities on Market1501 [2]. Each image corresponds to an metric-IFGSM [13] adversarial example and some randomly generated homogeneous modalities images. The same color means that they are obtained by transformation of the same image. Dots means adversarial examples.

Ablation studies. We studied the contributions of our channel fusion without augmentation (CF) , proactive defence (PD) and joint adversarial defence (JAD). From Table 5, we can see that defense accuracy increases with the increase of data diversity. As a passive defense, CS further significantly enhances defense performance. Table 6 shows that CS effectively reduces the adversarial effect and thus significantly enhances the defense with negligible performance degradation.

Comparison of state-of-the-arts. AT [12] needs to be customized according to attacks. Specifically, in order to defend an attack, it is necessary to add corresponding adversarial examples to train the model. In Table 7, the original accuracy (no attack) of defense methods based on AT [12] are the average accuracy of the defense models corresponding to the three attacks. AT+CS is the defense model combining the AT [12] and our CS, and it can be seen that the proactive defence and passive defence method CS (that is, our JAD) exert a better combined effect. This shows that our proactive and passive defenses can complement each other and have a better gaining effect. As with RRP [50], even if the attacker is aware of the existence of passive defenses, CS can still be effectively defended by a randomization mechanism that allows the resize to fluctuate within a certain range. Compared with AT [12], our JAD enhances Rank-1 by more than 14.5% in all white-box attacks.

In the tests of DMR [14] black-box attack, we use adversarial examples generated by Resnet50 [57] to attack the DenseNet [40] model. GOAT [13] model is training based on the adversarial samples generated online by the FNA attack [13] using triplet loss. GOAT+CS is the defense model combining the GOAT [13] and our CS. It can be seen from Table 8 that the defense accuracy of our JAD is far better than GOAT [13]. In addition, the experimental results show that the JAD is applicable to other baselines [59,60] and performs well. The strong baseline (SB) [59] is implemented based on the Resnet50 backbone network adding the batch normalization neck structure, and FastReID (FR) [60] is implemented based on the IBN-ResNet101 [62] backbone network. The method proposed in this paper has good defense effect in both white-box attack and black-box attack.



(a)Clean sample

(b)Adversarial sample

Figure 8. Comparison of Grad-CAM [63] activation map between normally trained model and our proactive defense model.

Visualization analysis. As the show in Figure 7, our proactive defense model with robust to color variations which is insensitive to the variations in the adversarial examples relative to the original examples. Therefore, we can observe that the features of adversarial example and homogeneous examples exhibit clustering effects.

Grad-CAM [63] uses the gradient information flowing into the last convolutional layer of the CNN to visualize the importance of each neuron in the output layer for the final prediction, by which it is possible to visualize which regions of the image have a significant impact on the prediction of a model. As shown in Figure 8*b*, we can see that the adversarial example successfully distracts the attention of the normally trained model and activates the opposite parts, while the our proactive defense model is still effectively activating some important parts.

5. Conclusion

In this paper, we proposed a color attack method (LTA) based on the local transformation, and further proposed a joint adversarial defense method (JAD) based on the feature-invariance mechanism to enhance the adversarial robustness of ReID. Finally, we used different network structures and baselines under different attack modes to conduct comparative experiments to verify the effectiveness of proposed attack method and joint defense method. Our future goal is to further enhance the stability of the proactive defense model, because we experimentally observed that the limitations of proactive defense are regular. Therefore, we will try to consider cross-datasets when training the model, and update the parameters of the model when a model is improved on the original and cross-domain dataset.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant Nos. U1805263, 61976053, 61672157.

References

[1] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C.H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[2] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jing dong Wang, , and Qi Tian. Scalable person re-identification:a benchmark. In *ICCV*, 2015. 1, 5, 6, 7, 8

[3] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *IEEE Transactions on Information Forensics and Security*, 16:728–739, 2017. 1

[4] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *CVPR*, pages 2898–2907, 2021. 1

[5] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, and Shiguang Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In *CVPR*, pages 2014–2023, 2021. 1

[6] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *CVPR*, pages 1522–1531, 2021. 1

[7] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, and Xiaoyun Yang. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *CVPR*, pages 13334–13343, 2021. 1

[8] Lu Pang, Yaowei Wang, YiZhe Song, Tiejun Huang, and Yonghong Tian. Cross-domain adversarial feature learning for sketch re-identification. 1

[9] Yi Li, Timothy M. Hospedales, Yizhe Song, and Shaogang Gong. Fine-grained sketch-based image retrieval by matching deformable part models. *BMVC*, pages 1–12, 2014. 1

[10] Lucas Beyer, Stefan Breuers, Vitaly Kurin, and Bastian Leibe. Towards a principled integration of multi-camera re-identification and tracking through optimal bayes filters. In *CVPRW*, 2017. 1

[11] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arXiv:1312.6199, 2014. 1

[12] Song Bai, Yingwei Li, Yuyin Zhou, Qizhu Li, and Philip H.S. Torr. Metric attack and defense for person re-identification. arXiv:1901.10650, 2019. 1, 2, 3, 6, 7, 8

[13] Quentin Bouniot, Romaric Audigier, and Angelique Loesch. Vulnerability of person re-identification models to metric adversarial attacks. In *CVPRW*, 2020. 1, 2, 3, 6, 7, 8

[14] Hongjun Wang, Guangrun Wang, Ya Li, Dongyu Zhang, and Liang Lin. Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking. In *CVPR*, 2020. 1, 2, 3, 7, 8

[15] Zhedong Zheng, Liang Zheng, Zhilan Hu, and Yi Yang. Open set adversarial examples. arXiv:1809.02681v1, 2018. 1, 3

[16] Zhibo Wang, Siyan Zheng, Mengkai Song, Qian Wang, Alireza Rahimpour, and Hairong Qi. advpattern: physical-world attacks on deep person re-identification via adversarially transformable patterns. In *ICCV*, 2019. 1, 2

[17] Yunpeng Gong. A general multi-modal data learning method for person re-identification. arXiv:2101.08533, 2021. 1, 3, 4, 5

[18] Mo Zhou, Zhenxing Niu, Le Wang, Qilin Zhang, and Gang Hua. Adversarial ranking attack and defense. In *ECCV*, 2020. 2, 3

[19] Giorgos Tolias, Filip Radenovic, and Ondrej Chum. Targeted mismatch adversarial attack: Query with a flower to retrieve the tower. In *ICCV*, 2019. 2

[20] Mo Zhou, Le Wang, Zhenxing Niu, Qilin Zhang, Yinghui Xu, Nanning Zheng, and Gang Hua. Practical relative order attack in deep ranking. In *ICCV*, 2021. 2

[21] Xiaodan Li, Jinfeng Li, Yuefeng Chen, Shaokai Ye, Yuan He, Shuhui Wang, Hang Su, and Hui Xue. Qair: Practical query-efficient black-box attacks for image retrieval. In *CVPR*, 2021. 2

[22] Jie Li, Rongrong Ji, Hong Liu, Xiaopeng Hong, Yue Gao, and Qi Tian. Universal perturbation attack against image retrieval. In *ICCV*, 2019. 2

[23] Mo Zhou and Vishal M. Patel. Enhancing adversarial robustness for deep metric learning. arXiv:2203.01439, 2022. 2, 3

[24] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019. 2

[25] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018. 2, 3

[26] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018. 2

[27] Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Improving the generalization of adversarial training with domain adaptation. In *ICLR*, 2019. 2

[28] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *ICLR*, 2019. 2, 3

[29] Tianyuan Zhang and Zhanxing Zhu. Interpreting adversarially trained convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 7502–7511, 2019. 2

[30] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *ICLR*, 2017. 2

[31] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning:from phenomena to black-box attacks using adversarial samples. arXiv:1605.07277, 2016. 2

[32] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2014. 3

[33] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018. 3

[34] Qiqi Xiao, Hao Luo, and Chi Zhang. Margin sample mining loss: A deep learning based method for person re-identification. arXiv:1710.00478, 2017. 3

[35] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. arXiv:1703.07737, 2017. 3

[36] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016. 3

[37] Uyeong Jang, Xi Wu, and Somesh Jha. Objective metrics and gradient descent algorithms for adversarial examples in machine learning. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, pages 262–277, 2017. 3

[38] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *IEEE Symposium on Security and Privacy*, pages 39–57, 2017. 3

[39] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person re-identification: Past, present and future. arXiv:1610.02984, 2016. 3

[40] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 3, 6, 8

[41] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *ICCV*, 2017. 3

[42] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. arXiv:1711.08184, 2017. 3

[43] 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 *ECCV*, 2018. 3

[44] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 3

[45] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 3, 6, 7

[46] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. In *ECCV*, 2018. 3

[47] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, 2018. 3

[48] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maatenn. Countering adversarial images using input transformations. In *ICLR*, 2018. 3

[49] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. A study of the effect of jpg compression on adversarial images. arXiv preprint arXiv:1608.00853, 2016. 3

[50] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *ICLR*, 2018. 3, 5, 6, 8

[51] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083, 2017. 3

[52] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L. Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *CVPR*, pages 501–509, 2019. 3

[53] Nicholas Carlini, Guy Katz, Clark Barrett, and David L. Dill. Ground-truth adversarial examples. In *ICLR*, 2018. 3

[54] Papernot N, Faghri F, Carlini N, Goodfellow I, Feinman R, Kurakin A, and et al. Technical report on the cleverhans v2.1.0 adversarial examples library. arXiv:1610.00768v6, 2016. 3

[55] Mang Ye, Jianbing Shen, Senior Member, IEEE, and Ling Shao. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE Transactions on Information Forensics and Security*, 16:728–739, 2021. 4

[56] Thierry Blu, Philippe Thévenaz, and Michael Unser. Linear interpolation revitalized. *IEEE Transactions on Image Processing*, 13(5):710–719, 2004. 5

[57] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 8

[58] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 7

[59] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPRW*, 2019. 7, 8

[60] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: a pytorch toolbox for general instance re-identification. arXiv:2006.02631, 2020. 7, 8

[61] L Maaten and G Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, pages 2579–2605, 2008. 8

[62] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018. 8

[63] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 8