Eliminate Deviation with Deviation for Data Augmentation and a General Multi-modal Data Learning Method

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

One of the challenges of computer vision is that it needs to adapt to color deviations in changeable environments. Therefore, minimizing the adverse effects of color deviation on the prediction is one of the main goals of vision task. Current solutions focus on using generative models to augment training data to enhance the invariance of input variation. However, such methods often introduce new noise, which limits the gain from generated data. To this end, this paper proposes a strategy eliminate deviation with deviation, which is named Random Color Dropout (RCD). Our hypothesis is that if there are color deviation between the query image and the gallery image, the retrieval results of some examples will be better after ignoring the color information. Specifically, this strategy balances the weights between color features and color-independent features in the neural network by dropouting partial color information in the training data, so as to overcome the effect of color devitaion. The proposed RCD can be combined with various existing ReID models without changing the learning strategy, and can be applied to other computer vision fields, such as object detection. Experiments on several ReID baselines and three common large-scale datasets such as Market1501, DukeMTMC, and MSMT17 have verified the effectiveness of this method. Experiments on Cross-domain tests have shown that this strategy is significant eliminating the domain gap. Furthermore, in order to understand the working mechanism of RCD, we analyzed the effectiveness of this strategy from the perspective of classification, which reveals that it may be better to utilize many instead of all of color information in visual tasks with strong domain variations.

1. Introduction

Person re-identification (ReID) is to match 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. The challenge of this task is that images captured by different cameras often contain significant intra-class variation caused by variations in viewpoint, human pose changes, occlusions, and color deviation under variable camera conditions, etc. As a result, the appearance of the same pedestrian image with great changes, making intra-class (the same pedestrian) metric distance larger than inter-class (different pedestrians). ReID usually combines representation learning [11–16] with metric learning [17–23], and combines classification loss [11, 13, 16] with triplet loss [22, 23] in the training stage to optimize the neural network. In the inference stage, it is only necessary to use Cosine distance or Euclidean distance to measure the similarity between the query image and the gallery image, and then rank the gallery images according to the similarity, and finally use the re-ranking technique [24] to further refine the search results.

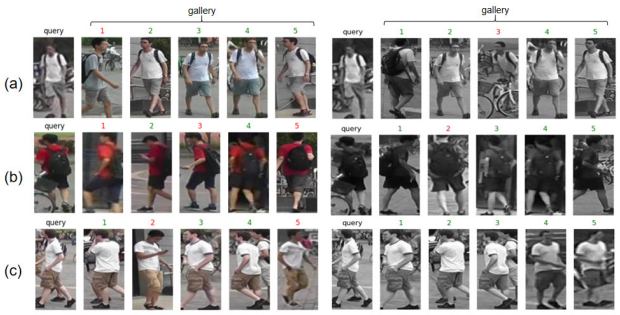

Figure 1. The retrieval results of the model trained with visible (RGB) image and the model trained with grayscale image on Market1501 [1] are displayed. It shows that the color deviation between the query image and gallery image will affect the retrieval results, and the retrieval results of some samples will be better after ignoring the color information. The numbers on the images indicate the rank of similarity in the retrieval results, the red and green numbers denote the wrong and correct results, respectively.

The complexity of the inherent challenge of ReID means that its demand for data has the same complexity, and the complex data demand is difficult to meet and balance by

the training set, which is also accompanied by the potential problem that the model overfits the only training data and lacks robustness. The dataset is hard to cover different camera environments and all their variations at different times, so the trained models tend to overfit the given training set and lack robustness to additional scenarios. It is no doubt that color features are important discriminative features, but color features instead limit the model to make correct predictions in some cases. For example, because white and gray, black and dark blue, and brown and yellow are similar under some lighting conditions, it is difficult for the model to make correct predictions for negative samples that are similar to target after overfitting the color deviation variations. As shown in (a) to (c) in Figure 1, the prediction results made by the model trained with grayscale images in this case are better after discarding the color bias.

Realistically, color deviations cause domain gaps exist both between datasets and within datasets [25, 26]. These color biases are practically inexhaustible. Instead of generating a variety of data to let the model "see" these variations (especially intra-class variations) [25] during training to enhance the robustness to input variations, it is better to balance the weight between color features and other important discriminant features implicitly. Aiming at the inherent color deviation problem that the images obtained under different shooting conditions, this paper proposes a strategy to eliminate deviation with deviation based on the assumption that the retrieval results of some samples will be better when discarding color information, which is named random color dropout (RCD). RCD balances the weight between color features and other important discriminant features in neural network by discarding part of the color information in the training data, so as to overcome the influence of color deviation. This strategy exists in various forms. For example, color deviation can be overcome with biased grayscale information or sketch information (or contour information). Taking grayscale as an example, it can be randomly selected a rectangular area in the RGB image and replace its pixels with the same rectangular area in the corresponding grayscale image , thus it generates a training image with different areas of biases during ReID model training. Compared with existing methods based on generative adversarial networks (GANs) [25–31], the proposed method is more lightweight and effective because it not only does not introduce new noise but also saves a large amount of computational resources. At the same time, this strategy enables the model to naturally have cross-modal retrieval [8, 9, 32–35] capabilities. For example, when taking the contour information as the intermediary to overcome the color deviation, the cross-modal retrieval between sketch and RGB visible image can be realized.

In addition, this paper analyzes the relationship between RCD and the generalization ability of neural networks from the perspective of classification, and reveals the intrinsic reasons that networks trained with RCD may outperform ordinary networks. Experiments show that the proposed method not only increases the robustness of the model to color deviation but also bridges the domain gap between different datasets, which has significant advantages over the existing state-of-the-art method. Taking the grayscale as an example, the RCD strategy proposed in this paper includes global grayscale transformation, local grayscale transformation, and a combination of these two. The method has the following advantages:

It is a lightweight approach which does not require any additional parameter learning or memory consumption. It can be combined with various CNN models without changing the learning strategy. It is a complementary approach to existing data augmentation.

The main contributions of this paper are summarized as follows:

• This paper proposes a learning strategy which against color deviation with information deviation, which decreases the overfitting and increases generalization ability of the model.

• A simple and effective cross-modal retrieval method is proposed, which does not need complex network design.

• This paper proves that the network trained with RCD may be better than the ordinary network from the perspective of classification.

• The strategy proposed in this paper is proved to be effective in improving ReID performance through extensive experiments and analysis. The effectiveness of the proposed method is verified on several baselines and representative datasets.

This work was previously published as a preprint on Arxiv and extended on the basis of it, including related demonstration and cross-modal retrieval.

2. Related Work

The complexity of the inherent challenge of ReID means that its demand for data has the same complexity, and the failure to fully meet the complex demand of the data is the source of the problem of overfitting and insufficient generalization of the model to the training data. Improving generalization ability is the focus of research in convolutional neural networks (CNNs). Therefore, data augmentation is effective in improving the generalization ability of the model.

2.1. Classic Data Augmentation

Many data augmentation [36] methods have been proposed, such as random cropping [36], flipping [37], which are well known to play an important role in classification, detection and ReID. CutMix [38] replaces one patch of an image with a patch from another image. Random erasing or

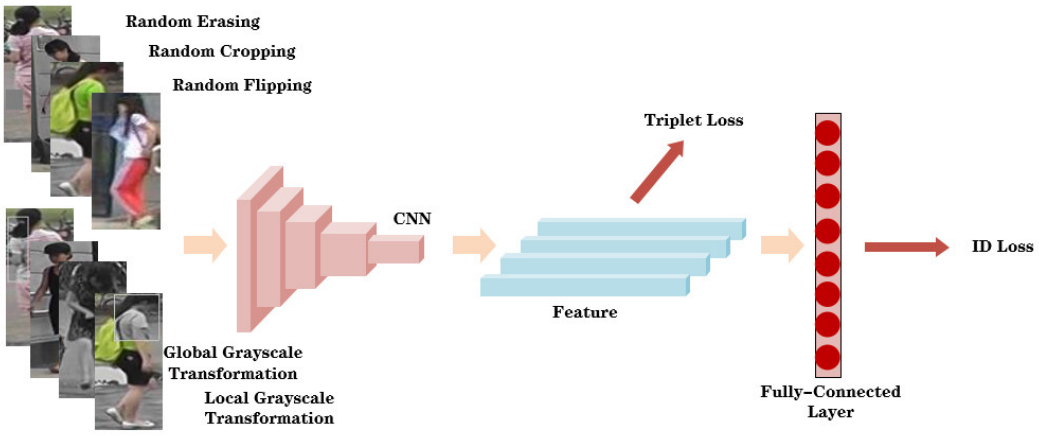

Figure 2. Framework diagram of our Random Color Dropout (RCD): The application of global grayscale transformation and local grayscale transformation in the framework.

cutout [39, 40] adds noise block to the image to regularize the network, while it helps to solve the occlusion problem in the ReID. The above methods are regarded as indispensable methods, and they are applied to various baselines [41–44]. These techniques have been proved to be effective in improving the prediction accuracy, and they are complementary to each other [41].

In solving the problem of color deviation, the early work [45] used the filter and the maximum grouping layer to learn the illumination transformation, divided the pedestrian image into more small pieces to calculate the similarity, and uniformly handled the problems of misalignment, occlusion and illumination variation under the deep neural network; [46] performed pre-processing before feature extraction and used multiscale Retinex algorithm to enhance the color information of light shaded regions to improve the color changes caused by lighting condition changes.

With the increasing maturity of GANs, GANs-based approaches for data augmentation have become an active research field.

2.2. Data Augmentation Based on GANs

The goal of these methods is to mitigate the effect of color deviation or human-pose variation, and to improve the robustness of the model by learning the invariant features from the variation of the input. The appearance details and the emphases generated by different GANs-based methods are also different, but their goal is all to compensate for the difference between the source and target domains. For example, CamStyle [52] generates new data for transferring different camera styles to learn invariant features between different cameras to increase the robustness of the model to camera style changes; CycleGAN [47] was applied in [29, 48] to transfer pedestrian image styles from one dataset to another; StarGAN [49] was used by [50] to generate pedestrian images with different camera styles. Wei et al. [26] proposed PTGAN to achieve pedestrian image transfer across different ReID datasets. It uses semantic segmentation to extract foreground masks to assist style transfer, and converts the background into the desired style of the dataset while keeping the foreground unchanged. different from global style transfer, DGNet [25] utilizes GANs to transfer clothing among different pedestrians by manipulating appearance and structural details to generate more diverse data to reduce the impact of color changes on the model, which effectively improves the generalization ability of the model. In addition, [51] uses 3D engine and environment rendering technology to build a virtual pedestrian data set with multiple lighting conditions, which is combined with other large real data sets to jointly pre-train a model.

The method proposed in this paper has been partially validated in other works. [52] proved that the lack of robustness to color deviation is one of the main reasons why the model is vulnerable to adversarial metric attacks [53–55], and enhanced the model's adversarial defense using the method proposed in this paper; It is adopted in the new baseline proposed in [56, 57] to help increase the generalization of the model. In addition, [58] showed that the method proposed in this paper is also suitable for object detection.

3. Proposed Methods

The RCD strategy proposed in this paper includes global transformation, local transformation, and a combination of the two. Taking grayscale as an example, it includes global grayscale transformation, local grayscale transformation, and combinations of the two. At the end of this subsection, we give the corresponding analysis of the proposed method. The framework of this method is showed in Figure 2.

Algorithm 1: Global Graycale Transformation

Input : Input image I;
Graycale transformation probability p;
.
Output: Grayscale images I^*.
Initialization: $p_1 \leftarrow$ Rand (0, 1).
if $p_1 \geq p$ **then**
  $I^* \leftarrow I$;
  return I^*.
else
  $I^* \leftarrow$ t(I);
  return I^*.
end

3.1. Global Grayscale Transformation

In the data loading, it randomly samples K identities and M images of per person to constitute a training batch which size equals to $B = K \times M$. The set is denoted as $x^v = \{x_i^v | i = 1, 2, ..., M \times K\}$, where $x_i^v = \{x_i^v | y_i\}$ represents the i-th sample image of the training batch, and y_i represents the class label of the pedestrian.

Taking the grayscale as an example, this method randomly performs global grayscale transformation on the training batch with a probability, and then inputs into the model for training. This process can be defined as:

$$I_g = t(R, G, B) \tag{1}$$

where $t(\bullet)$ is the grayscale image conversion function, which is implemented by performing pixel-by-pixel accumulation calculations on the R, G, and B channels of the original visible RGB image; y is the label of the sample, the converted grayscale image label of x^g are the same as the original ones:

$$(x^g | y) = (x^v | y) \tag{2}$$

the procedure of LGT is shown in Algorithm.1.

3.2. Local Grayscale Transformation

In addition to transforming the data globally, we also consider transforming the data locally so that the model adapt better to the significantly varying bias due to color dropout from the local variation.

The local grayscale transformation (LGT) for each visible image x^v can be achieved by the following equations:

$$x^g = t(x^v), \tag{3}$$

$$rect = RandPosition(x^v), \tag{4}$$

$$x^{lg} = LGT(x^v, x^g, rect) \tag{5}$$

Algorithm 2: Local Graycale Transformation

Input : Input image I;
Image size W and H;
Area of image S;
Transformation probability p;
Area ratio range s_l and s_h;
Aspect ratio range r_1 and r_2.
Output: Transformed image I^*.
Initialization: $p_1 \leftarrow$ Rand (0, 1).
if $p_1 \geq p$ **then**
  $I^* \leftarrow I$;
  return I^*.
else
  while *True* **do**
    $S_t \leftarrow$ Rand $(s_l, s_h) \times S$;
    $r_t \leftarrow$ Rand (r_1, r_2);
    $H_t \leftarrow \sqrt{S_t \times r_t}$, $W_t \leftarrow \sqrt{\frac{S_t}{r_t}}$;
    $x_t \leftarrow$ Rand $(0, W)$, $y_t \leftarrow$ Rand $(0, H)$;
    if $x_t + W_t \leq W$ *and* $y_t + H_t \leq H$ **then**
      $Position \leftarrow (x_t, y_t, x_t + W_t, y_t + H_t)$;
      $I(Position) \leftarrow t(Position)$;
      $I^* \leftarrow I$;
      return I^*.
    end
  end
end

and

$$(x^{lg} | y) = (x^v | y) \tag{6}$$

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

In the process of model training, we conduct LGT randomly transformation on the training batch with a probability. For an image I in a batch, denote the probability of it undergoing LGT be p_r, and the probability of it being kept unchanged be $1 - p_r$. In this process, it randomly selects a rectangular region in the image and replaces it with the pixels of the same rectangular region in the corresponding grayscale image. Thus, training images which include regions with different levels of grayscale are generated. Among them, s_l and s_h are the minimum and maximum values of the ratio of the image to the randomly generated rectangle area, and the S_t of the rectangle area limited between the minimum and maximum ratio is obtained by $S_t \leftarrow Rand(s_l, s_h) \times S$, r_t is a coefficient used to determine the shape of the rectangle. It is limited to the interval $(r_1, r_2$

). x_t and y_t are randomly generated by coordinates of the upper left corner of the rectangle. If the coordinates of the rectangle exceed the scope of the image, the area and position coordinates of the rectangle are re-determined. When a rectangle that meets the above requirements is found, the pixel values of the selected region are replaced by the corresponding rectangular region on the grayscale image converted from RGB image. As a result, training images which include regions with different levels of grayscale are generated, and the object structure is not damaged. The procedure of LGT is shown in Algorithm.2.

3.3. Loss function

ReD usually combines classification loss and triplet loss to train the model [59, 60]. Here, we use x_i^v to denote the i-th RGB image in a training batch, and x_i^g to denote the image obtained after GGT or LGT conversion. Thenthe features of x_i^v and x_i^g can be expressed as:

$$\begin{cases} f_i^v = f(x_i^v) \\ f_i^g = f(f_i^g) \end{cases} \tag{7}$$

The Euclidean distance between two samples x_i^g and x_j^g is denoted as $D(x_i^g, x_j^g)$, where The subscript $\{i, j\}$ denotes the image index in the training batch. Formally, let x_i^g be the anchor sample, the triple $\{x_i^g, x_j^g, x_k^g\}$ is selected in the following way:

$$P_{i,j}^g = \max_{\forall y_i = y_j} D(x_i^g, x_j^g) \tag{8}$$

$$N_{i,k}^g = \min_{\forall y_i \neq y_j} D(x_i^g, x_k^g) \tag{9}$$

For each anchor point x_i^g, the above strategy selects the positive sample pair with the same pedestrian class label and the farthest the most distant positive sample pair with the same pedestrian category label and the nearest negative sample pairs, forming a triplet $\{x_i^g, x_j^{g^+}, x_k^{g^-}\}$ for mining grayscale information. In general, the use of The boundary parameter ε is used to control the spacing of the positive and negative sample pairs. In summary, we can define the following triadic loss for training:

$$L_g = \frac{1}{n} \sum_{i=1}^{n} \max[\varepsilon + D(x_i^g, x_j^{g^+}) + D(x_i^g, x_k^{g^-})] \tag{10}$$

In addition, x^v and x^g are trained using a shared identity classifier ϕ. The predicted probability of identity labels y_i is define as $p(y_i|x_i^g; \phi)$. The ID loss is represented as follows:

$$L_{ide} = -\frac{1}{n} \sum_{i=1}^{n} log(p(y_i|x_i^g; \phi)) \tag{11}$$

Therefore, the overall loss during random grayscale transformation is:

$$L_{total} = L_g + L_{ide} \tag{12}$$

3.4. Analysis of Random Color Dropping Policy

Here suppose there are m instances, the expected output, i.e. $D = [d_1, d_2, \ldots, d_m]^T$ where d_j denotes the expected output on the j-th instance, and the actual output of the i-th component neural network, i.e. $F_i = [f_{i1}, f_{i2}, \ldots, f_{im}]^T$ where f_{ij} denotes the actual output of the i-th component network on the j-th instance. D and F_i satisfy that $d_j \in \{-1, +1\}(j = 1, 2, \ldots, m)$ and $f_{ij} \in \{-1, +1\}(i = 1, 2, \ldots, N; j = 1, 2, \ldots, m)$ respectively. It is obvious that if the actual output of the i-th component network on the j-th instance is correct according to the expected output then $f_{ij}d_j = +1$, otherwise $f_{ij}d_j = -1$. Thus the generalization error of the i-th component neural network on those m instances is:

$$E_i = \frac{1}{m} \sum_{j=1}^{m} Error(f_{ij}d_j) \tag{13}$$

where $Error(x)$ is a function defined as:

$$Error(x) = \begin{cases} 1, & if \quad x = -1 \\ 0.5, & if \quad x = 0 \\ 0, & if \quad x = 1 \end{cases} \tag{14}$$

Here we introduce a vector $Sum = [Sum_1, Sum_2, \ldots, Sum_m]^T$ where Sum_j denotes the sum of the actual output of all the component neural networks on the j-th instance, i.e.

$$Sum_j = \sum_{i=1}^{N} f_{ij} \tag{15}$$

Then the output of the neural network ensemble on the j-th instance is:

$$\hat{f}_j = Sgn(Sum_j) \tag{16}$$

where $Sgn(x)$ is a function defined as:

$$Sgn(x) = \begin{cases} 1, & if \quad x > 0 \\ 0, & if \quad x = 0 \\ -1, & if \quad x < 0 \end{cases} \tag{17}$$

It is obvious that $\hat{f}_j \in \{-1, 0, +1\}(j = 1, 2, \ldots, m)$. If the actual output of the ensemble on the j-th instance is correct according to the expected output then $\hat{f}_j d_j = +1$; if it is wrong then $\hat{f}_j d_j = -1$; otherwise $\hat{f}_j d_j = 0$, which means that there is a tie on the j-th instance, e.g. three component networks vote for +1 while other three networks vote for -1. Thus the generalization error of the ensemble is:

$$\hat{E} = \frac{1}{m} \sum_{j=1}^{m} Error(\hat{f}_j d_j) \tag{18}$$

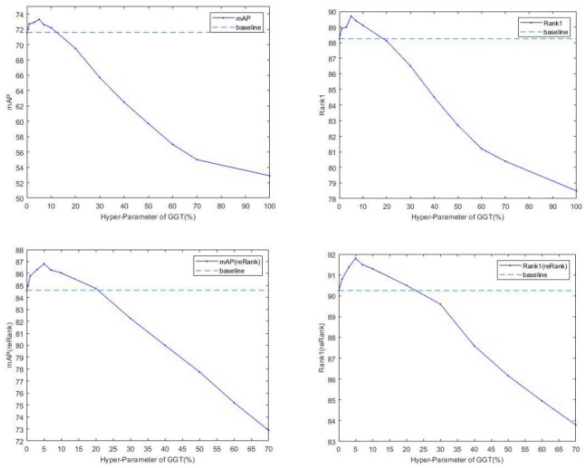

Figure 3. Performance of GGT under different hyperparameters on Market1501.

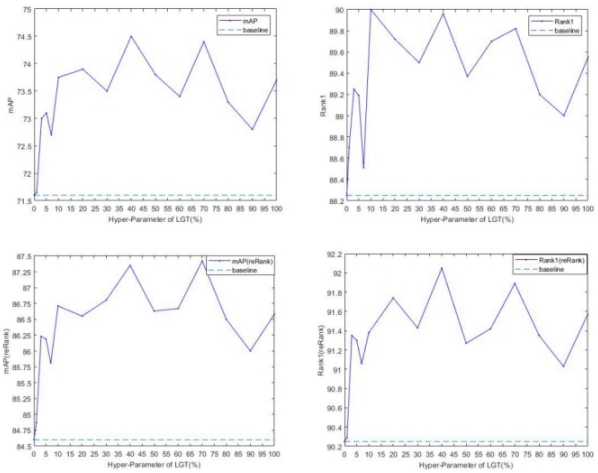

Figure 4. Performance of LGT under different hyperparameters on Market1501.

Here suppose that the k-th component neural network is trained using grayscale images. Then the output of the new set on the j-th instance is:

$$\hat{f}_j' = Sgn(Sum_{j(j\neq k)} - f_{kj}) \tag{19}$$

and the generalization error of the new ensemble is:

$$\hat{E}' = \frac{1}{m}\sum_{j=1}^{m} Error(\hat{f}_j' d_j) \tag{20}$$

It is assumed that a certain number of networks with deviations will not affect the performance of the overall neural network, and the retrieval results of some examples will be better after ignoring the color information.

$$Error(\hat{f}_j' d_j) \leqslant Error(\hat{f}_j d_j) \tag{21}$$

From Eq.(18) and Eq.(20) we can derive that if Eq.(21) is satisfied then $\hat{E}$ is not smaller than $\hat{E}'$, which means that the ensemble including k-th component neural network which is trained using grayscale images is better than the one no including:

$$\sum_{j=1}^{m}\{Error(Sgn(Sum_j)d_j) - Error(Sgn(Sum_{j(j\neq k)+fkj})d_j)\} \geq 0 \tag{22}$$

4. Comparison and Analysis

4.1. Datasets and Evaluation criteria

Datasets. Market-1501 [1] includes 1,501 pedestrians captured by six cameras (five HD cameras and one low-definition camera). DukeMTMC [61] 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. MSMT17 [26], created in winter, was presented in 2018 as a new, larger dataset closer to real-life scenes, containing a total of 4,101 individuals and covering multiple scenes and time periods.

These three datasets are currently the largest datasets of ReID, and they are also the most representative because they collectively contain multi-season, multi-time, HD, and low-definition cameras with rich scenes and backgrounds as well as complex lighting variations.

Sketch ReID dataset [8] contains 200 persons, each of which has one sketch and two photos. Photos of each person were captured during daytime by two cross-view cameras. It cropped the raw images (or video frames) manually to make sure that every photo contains the one specific person. It have a total of 5 artists to draw all persons'sketches and every artist has his own painting style.

Evaluation criteria. Following existing works [1], 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.

4.2. Hyper-Parameter Setting

During CNN training, two hyper-parameters need to be evaluated. One of them is GGT probability p. Firstly, we take the hyper-parameter p as 0.01, 0.03, 0.05, 0.07, 0.1, 0.2, 0.3,..., 1 for the GGT experiments. Then we take the value of each parameter for three independent repetitions of the experiments. Finally, we calculate the average of the final result. The results of different p are shown in Fig.3. We can see that when $p = 0.05$, the performance of the

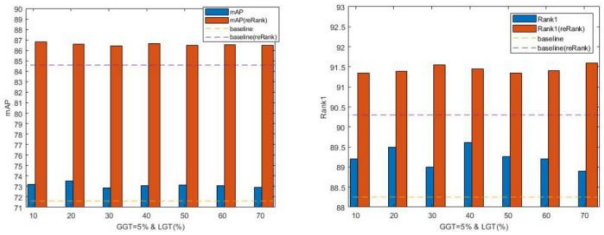

Figure 5. Performance of combining GGT with LGT under different hyperparameters on Market1501.

model reaches the maximum value in Rank-1 and mAP. If we do not specify, the hyper- parameter is set $p = 0.05$ in the next experiments.

Another hyper-parameter is LGT probability p_r. We take the hyper-parameter p_r as the same as above for the LGT experiments, whose selection process is similar to the above p. The results of different p_r are shown in Fig.4.

Obviously, when $p_r = 0.4$ or $p_r = 0.7$, the model achieves better performance. And the best performance is achieved when $p_r = 0.4$. If we do not specify, the hyper-parameter is set $p_r = 0.4$ in the subsequent experiments.

Evaluation of GGT and LGT. Compared with the best results of GGT on baseline [44], the accuracy of LGT is improved by 0.5% and 1.4% on Rank-1 and mAP, respectively. Under the same conditions using re-Ranking [24], the accuracy of Rank-1 and mAP is improved by 0% and 0.4%, respectively. Therefore, the advantages of LGT are more obvious when re-Ranking is not used. However, Fig.4 also shows that the performance improvement brought by LGT is not stable enough because of the obvious fluctuation in LGT, while the performance improvement brought by GGT is very stable. Therefore, we improve the stability of the method by combining GGT with LGT.

Evaluation by Combining GGT with LGT. First, we fix the hyper-parameter value of GGT to $p = 0.05$, then keep the control variable unchanged to further determine the hyper-parameter of LGT. Finally, we take the hyper-parameter pr of LGT to be 0.1, 0.2, ···, 0.7 to conduct combination experiments of GGT and LGT, and conduct 3 independent repeated experiments for each parameter p_r to get the average value. The result is shown in 5. It can be seen that the performance improvement brought by the combination of GGT and LGT is more stable and with less fluctuation, and the comprehensive performance of the model is the best when the hyper-parameter value of LGT is $p_r = 0.4$.

4.3. Comparison Experiments

Performance comparison and analysis. We first evaluate baseline [44] on the Market-1501 dataset [1]. To be consistent with recent works, we follow the new training/testing protocol to conduct our experiments by k-reciprocal re-ranking (RK) [24]. As can be seen from Fig.3 and Fig.4, our method improves by 1.2% on Rank-1 and 3.3% on mAP on

Figure 6. Diagram of Global Sketch Transformation (GST) and Local Sketch Transformation (LST).

Table 1. Performance comparison on Market1501 dataset.

Methods	Market1501	
	Rank-1(%)	mAP(%)
IANet [62]	94.4	83.1
DGNet [25]	94.8	86.0
SCAL [63]	95.8	89.3
Circle Loss [64]	96.1	87.4
SB [59]	94.5	85.9
SB [59] + RK [24]	95.4	94.2
SB + GGT(ours)	**94.6**	85.7
SB + GGT+ RK(ours)	**96.2**	**94.7**
SB + LGT(ours)	**95.1**	**87.2**
SB + LGT + RK(ours)	**95.9**	**94.4**
FastReID [60]	96. 3	90.3
FastReID + RK	96.8	95.3
FastReID + GGT(ours)	**96.5**	**91.2**
FastReID + GGT + RK(ours)	**96.9**	**95.6**

the baseline, and 1.5% on Rank-1 and 2.1% on mAP above baseline in the same conditions using the re-Ranking [24].

Secondly, we further test the method in this paper on the baselines [59, 60] with better performance. As we can see from Table.1 to Table.3, the best results of our method improve by 0.6% and 1.3% on the Rank-1 and mAP on the strong baseline [59], respectively, and 0.8% and 0.5% Rank-1 and mAP above baseline under the same conditions using the re-Ranking [24], respectively. On fastReID [60], our method is 0.2% higher and 0.9% than baseline in Rank-1 and mAP, respectively, and higher 0.1% and 0.3% than baseline under using re-Ranking.

The default configuration on the Strong Baseline [59] and FastReID [60] uses data augmentation such as random flipping [37], cropping [36], and erasing [39]. The method proposed in this paper further improves the model accuracy on the basis of using them, which shows that our method can be combined with other data augmentation methods.

Table 2. Performance comparison on DukeMTMC dataset.

Methods	DukeMTMC	
	Rank-1(%)	mAP(%)
IANet [62]	87.1	73.4
DGNet [25]	86.6	74.8
SCAL [63]	89.0	79.6
SB [59]	86.4	76.4
SB + RK [24]	90.3	89.1
SB + GGT(ours)	**87.8**	**77.3**
SB + GGT+ RK(ours)	**90.9**	**89.2**
SB + LGT(ours)	**87.3**	**77.3**
SB + LGT + RK(ours)	**91**	**89.4**
FastReID [60]	92.4	83.2
FastReID + RK	94.4	92.2
FastReID + LGT(ours)	**92.8**	**84.2**
FastReID + LGT + RK(ours)	**94.3**	**92.7**

Table 3. Performance comparison on MSMT17 dataset.

Methods	MSMT17	
	Rank-1(%)	mAP(%)
IANet [62]	75.5	46.8
DGNet [25]	77.2	52.3
RGA-SC [65]	80.3	57.5
SCSN [66]	83.8	58.5
AdaptiveReID [67]	81.7	62.2
FastReID [60]	85.1	63.3
FastReID + GGT(ours)	**86.2**	**65.3**
FastReID + GGT&LGT(ours)	**86.2**	**65.9**

Cross-domain tests. Cross-domain person re-identification aims at adapting the model trained on a labeled source domain dataset to another target domain dataset without any annotation. It is pointed out by [59] that the higher accuracy of the model does not mean that it has better generalization capacity. In response to the above potential problems, we use cross-domain 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 [1] and DukeMTMC [61]. In order to further explore the effectiveness of the proposed method in cross-domain experiments, we use GGT to conduct the following cross-domain experiments on strong baseline [59]. The experiments are shown in Table.4.

In Table 4, +REA means that the trick of Random Erasing is used in model training, -REA means turning it off. Experimental results show that random erasing [39] can also significantly improve the performance of the ReID model, but it will cause a significant drop in cross-domain performance. The proposed method can not only significantly improve the cross-domain performance of the ReID model, but also be more robust because of learning more discriminative features.

Table 4. The performance of different models is evaluated on cross-domain dataset. M→D means that we train the model on Market1501 [1] and evaluate it on DukeMTMC [61].

Methods	Cross-Domain	
	M→D	D→M
	Rank-1/mAP(%)	Rank-1/mAP(%)
SB [59]+REA [39]+RK	33.6/24.3	51.6/32.3
SB+REA+GGT+RK(ours)	**37.8/27.8**	**55.4/35.7**
SB-REA+RK	45.5/37.0	58.2/37.8
SB-REA+GGT+RK(ours)	**48.2/37.9**	**65.0/43.7**

Table 5. Performance comparison between our LGT conversion and DGNet data augmentation on Market1501.

Methods	Market1501	
	Rank-1	mAP(%)
Baseline [44]	88.8	71.6
Baseline + DGNet [25]	88.9	72.1
Baseline+ LGT(ours)	**90.0**	**74.9**

Table 6. Cross-domain performance comparison between our LGT and DGNet on Market1501.

Methods	Market1501→DukeMTMC	
	Rank-1	mAP
Baseline [44]	37.8	27.0
Baseline + DGNet [25]	36.7	25.6
Baseline+ LGT(ours)	**39.7**	**27.9**

Table 7. Performance comparison on Market1501 dataset.

Methods	Market1501	
	Rank-1(%)	mAP(%)
Baseline [44]	88.8	71.6
Baseline + RK [24]	**90.5**	**85.2**
Baseline+ GST+LST(ours)	**88.9**	**72.6**
Baseline+ GST+LST+RK(ours)	**91.2**	**86.8**

Table 8. Performance comparison between our RCD and Adversarial Feature Learning on Sketch ReID dataset.

Sketch ReID dataset	Rnak-1(%)	Rnak-5(%)	Rank-10(%)
AFL [8]	34.0	56.3	72.5
GST+LST(ours)	**42.5**	**70.0**	**87.5**

Comparison of state-of-the-arts. A comparison of the performance of our method with the state-of-the-art methods DGNet [25] on Market1501 [1] is shown in Table.5 and Table.6. As can be seen from Table.5, our method delivers a performance improvement that far exceeds that of DGNet,

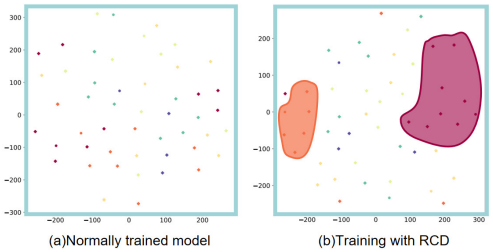

Figure 7. t-SNE [68] visualization of six randomly selected images with different identities on Market1501 [1]. Each image corresponds to the randomly generated images with color deviation. The same color means that they are obtained by transformation of the same image. Dots means original example.

the state-of-the-art GAN-based method, by more than 2.7 percentage points on mAP, which suggests that the proposed method is superior to existing data augmentation.

As can be seen from Table.6, the generalization ability of the proposed method in cross-domain tests is improved by 1.9 percentage points in the Rank-1 compared with the baseline [44], which further shows that the proposed method is better than the existing data augmentation based on generative models. It is worth noting that when the data generated by DGNet is used for model training, the cross-domain performance of the model is poorly, which confirms the point of this paper that color deviation is is difficult to exhaust and that instead of enhancing robustness to input changes by generating a variety of data for the model to "see" during training, it is better to implicitly reduce the weight of the model in the discriminant feature of color information.

Cross-modal retrival. Another form of strategy proposed in this paper is to take sketch image as the intermediary of balancing weight. By applying the proposed global homogeneity transformation and local homogeneity transformation, the sketch image is transformed as a homogeneous image, as shown in Figure.6. It can not only improve the robustness of the model, but also realize the sketch-based ReID. We can see this clearly from Table..7 and Table..8.

In terms of cross-modal retrieval [8, 69, 70], in order to match images of different modalities, existing approaches usually achieve cross-modal retrieval with the help of attention mechanisms [69], multi-stream networks [70], and generative adversarial networks [8]. Lu et al. proposed a cross-domain adversarial feature learning (AFL) method for sketch re-identification and contributed the sketch character re-identification dataset [8].

In order to make a fair comparison, as same as AFL [8], the method proposed in this paper is firstly trained on the Market-1501 dataset, and then fine tuned on sketch ReID dataset. In parameter setting, this paper set 5% Global Sketch Transformation and 70% Local Sketch Transformation. The experiment result shows that the performance improvement in the Sketch Re-identification more than 8%. This experiment also shows the generality of the proposed method.

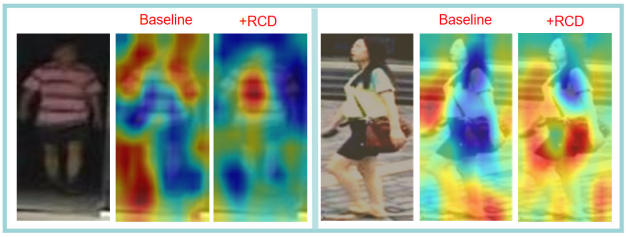

Figure 8. Comparison of Grad-CAM [71] activation map between normally trained model and our proactive defense model.

Visualization analysis. As the show in Figure 7, the model trained by DCR is robust to color variations. Therefore, we can observe that the features of examples with color deviation exhibit clustering effects better.

Grad-CAM [71] 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, we can see that the the normally trained model activates irrelevant parts in the case of severe color deviation, while the model trianed with RCD is still effectively activating some important parts.

5. Conclusion

In this paper, a simple, effective and general strategy that can be applied in computer vision to overcome color deviation. Neither does the method require large scale training like GAN, nor introduces any noise. The method uses a random homogeneous transformation to realize the modeling of different modal relationships. The model balances the weights between color features and discriminative non-color features by fitting differentiated homogeneous information in a mixed domain with information bias during the training process, thus reducing the negative impact of color deviation on ReID. In addition, this paper reveals the intrinsic reasons why networks trained with RCD outperform ordinary networks from a classification perspective. At the same time, experiments on several datasets and baselines show that the proposed method is effective and outperforms the state of the arts, and extra experiments show that the proposed strategy has natural cross modal properties.

References

[1] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 1, 6, 7, 8, 9

[2] 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

[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, 2, 6, 8, 9

[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, 2

[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] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 14(1):1–20, 2017. 1

[12] Tetsu Matsukawa and Einoshin Suzuki. Person re-identification using cnn features learned from combination of attributes. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 2428–2433. IEEE, 2016. 1

[13] Xing Fan, Wei Jiang, Hao Luo, and Mengjuan Fei. Spherereid: Deep hypersphere manifold embedding for person re-identification. *Journal of Visual Communication and Image Representation*, 60:51–58, 2019. 1

[14] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95:151–161, 2019. 1

[15] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[16] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, 28(6):2860–2871, 2019. 1

[17] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, and Neil M Robertson. Ranked list loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5207–5216, 2019. 1

[18] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European conference on computer vision*, pages 791–808. Springer, 2016. 1

[19] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *European conference on computer vision*, pages 135–153. Springer, 2016. 1

[20] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1

[21] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26(7):3492–3506, 2017. 1

[22] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the iEEE conference on computer vision and pattern recognition*, pages 1335–1344, 2016. 1

[23] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1

[24] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 1, 7, 8

[25] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2138–2147, 2019. 2, 3, 7, 8

[26] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 2, 3, 6

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[28] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 5157–5166, 2018. 2

[29] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 994–1003, 2018. 2, 3

[30] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4099–4108, 2018. 2

[31] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–667, 2018. 2

[32] Yuanxin Zhu, Zhao Yang, Li Wang, Sai Zhao, Xiao Hu, and Dapeng Tao. Hetero-center loss for cross-modality person re-identification. *Neurocomputing*, 386:97–109, 2020. 2

[33] Zhipeng Huang, Jiawei Liu, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification. *arXiv preprint arXiv:2203.01735*, 2022. 2

[34] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, volume 1, page 2, 2018. 2

[35] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4610–4617, 2020. 2

[36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2, 7

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 7

[38] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 2

[39] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020. 3, 7, 8

[40] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3

[41] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 3

[42] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 3

[43] Zhedong Zheng, Tao Ruan, Yunchao Wei, Yi Yang, and Tao Mei. Vehiclenet: Learning robust visual representation for vehicle re-identification. *IEEE Transaction on Multimedia (TMM)*, 2020. 3

[44] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):13, 2018. 3, 7, 8, 9

[45] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014. 3

[46] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015. 3

[47] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3

[48] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 598–607, 2019. 3

[49] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 3

[50] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–188, 2018. 3

[51] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 189–205, 2018. 3

[52] Yunpeng Gong, Liqing Huang, and Lifei Chen. Person re-identification method based on color attack and joint defence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4313–4322, June 2022. 3

[53] Quentin Bouniot, Romaric Audigier, and Angelique Loesch. Vulnerability of person re-identification models to metric adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 794–795, 2020. 3

[54] Song Bai, Yingwei Li, Yuyin Zhou, Qizhu Li, and Philip HS Torr. Metric attack and defense for person re-identification. 2019. 3

[55] Hongjun Wang, Guangrun Wang, Ya Li, Dongyu Zhang, and Liang Lin. Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep misranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 342–351, 2020. 3

[56] Xingyang Ni and Esa Rahtu. Flipreid: Closing the gap between training and inference in person re-identification. In *2021 9th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6. IEEE, 2021. 3

[57] Minghui Chen, Zhiqiang Wang, and Feng Zheng. Benchmarks for corruption invariant person re-identification. *arXiv preprint arXiv:2111.00880*, 2021. 3

[58] Seong-Eun Ryu and Kyung-Yong Chung. Detection model of occluded object based on yolo using hard-example mining and augmentation policy optimization. *Applied Sciences*, 11(15):7093, 2021. 3

[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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 5, 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. 5, 7, 8

[61] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*, pages 3754–3762, 2017. 6, 8

[62] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8514–8522, 2019. 7, 8

[63] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 371–381, 2019. 7, 8

[64] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020. 7

[65] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 3186–3195, 2020. 8

[66] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. Salience-guided cascaded suppression network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3300–3310, 2020. 8

[67] Xingyang Ni, Liang Fang, and Heikki Huttunen. Adaptive l2 regularization in person re-identification. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9601–9607. IEEE, 2021. 8

[68] L Maaten and G Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, pages 2579–2605, 2008. 9

[69] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *Proceedings of the IEEE international conference on computer vision*, pages 5551–5560, 2017. 9

[70] Emrah Basaran, Muhittin Gökmen, and Mustafa E Kamasak. An efficient framework for visible–infrared cross modality person re-identification. *Signal Processing: Image Communication*, 87:115933, 2020. 9

[71] 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. 9