Attack-Guided Perceptual Data Generation for Real-world Re-Identification

Yukun Huang* Xueyang Fu* Zheng-Jun Zha†
University of Science and Technology of China, China
kevinh@mail.ustc.edu.cn, {xyfu,zhazj}@ustc.edu.cn

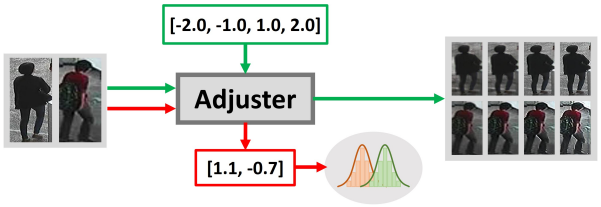
Abstract

In unconstrained real-world surveillance scenarios, person re-identification (Re-ID) models usually suffer from different low-level perceptual variations, e.g.*, cross-resolution and insufficient lighting. Due to the limited variation range of training data, existing models are difficult to generalize to scenes with unknown perceptual interference types. To address the above problem, in this paper, we propose two disjoint data-generation ways to complement existing training samples to improve the robustness of Re-ID models. Firstly, considering the sparsity and imbalance of samples in the perceptual space, a dense resampling method from the estimated perceptual distribution is performed. Secondly, to dig more representative generated samples for identity representation learning, we introduce a graph-based white-box attacker to guide the data generation process with intra-batch ranking and discriminate attention. In addition, two synthetic-to-real feature constraints are introduced into the Re-ID training to prevent the generated data from bringing domain bias. Our method is effective, easy-to-implement, and independent of the specific network architecture. Applying our approach to a ResNet-50 baseline can already achieve competitive results, surpassing state-of-the-art methods by +1.2% at Rank-1 on the MLR-CUHK03 dataset.*

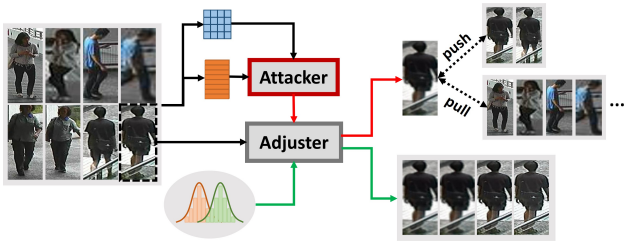
1. Introduction

Person re-identification (Re-ID) aims to identify the sample person across non-overlapping cameras, which can be viewed as a sub-task of image retrieval. However, due to identity-unrelated drastic variations, learning a robust and discriminative identity representation for real-world Re-ID in unconstrained scenarios is challenging. In general, these variations can be divided roughly into two categories: low-level **perceptual variations** (referred to as visual degradation), such as resolution and illumination; and high-level **semantic variations**, including view, pose, occlusions, *etc*.

*Co-first authors.
†Corresponding author.



(a) A lightweight image adjuster which can predict the resolution quality score of input images (Red Line) and adjust the image resolutions based on given values (Green Line).



(b) Taking the overall resolution distribution of the data source as prior knowledge, global-aware dense resolution augmentation can be performed for each sample (Green Line). An intra-batch white-box attacker is further introduced to provide guidance from high-level vision tasks (Red Line).

Figure 1. An overview of proposed **G**lobal-**A**ware and **A**ttack-**G**uided perceptual data generation (**GAAG**).

Compared with the former, semantic variations have been explored sufficiently by existing Re-ID methods, *e.g.*, predefined regional partition [36, 35, 30] and human part alignment [18, 51]. In this work, we focus on low-level perceptual variations and take cross-resolution Re-ID as the main task.

Due to the powerful representation learning capability, deep convolutional neural networks-based Re-ID models [26, 43, 29] can effectively deal with these variations in constrained scenarios. However, since real-world applications are more diverse and unpredictable, these deep models that rely on training data heavily are difficult to generalize to unseen situations. Although collecting enough labeled data is a feasible solution, it is too expensive and impractical to build a manually labeled database covering all possible situations. Therefore, many works attempt to complement the

training data by adjusting the original samples [25, 48] or synthesizing new ones [47, 28, 46].

Conventional augmentations, *e.g*., random crop and random horizontal flip, have been widely used for the Re-ID task. Thus we mainly discuss data augmentation methods based on synthetic strategies and classify them into two types: **(1) Engine-based generation.** Based on 3D rendering engines, controllable person generation [2, 1, 34] with different poses, backgrounds, *etc*., can be realized. Although these methods promote quantitative analysis of how visual factors influence the Re-ID system, artificial images are unsuitable as training data due to significant differences in style and appearance with real-world data. **(2) GAN-based generation.** Generative Adversarial Networks (GANs) [11] are also widely used for data generation. To our best knowledge, Zheng *et al*. [47] firstly introduce GANs into Re-ID for unlabeled data generation. By interpolating or swapping the disentangled intermediate features, [28, 46] achieve sufficiently realistic image generation. However, due to the limited number of original samples that provide prior information, the synthesized images lack sufficient diversity.

In this paper, to alleviate the interference of low-level perceptual variations in real-world surveillance scenarios, we propose a new **G**lobal-**A**ware and **A**ttack-**G**uided perceptual data generation approach (**GAAG**) by combining disentangled image generation and adversarial attack.

Specifically, we design a lightweight disentangled generative model (denoted as *Adjuster* in Figure 1) with two functions to predict the perceptual quality score of an input image and adjust the input image based on a given score. The first function aims to estimate the overall distribution of perceptual variations in the data source, while the second one takes the estimated distribution as prior knowledge to perform global-aware dense augmentation on each training instance. In this way, the perceptual diversity of training data can be enriched without changing identity semantics. However, it is not optimal to directly use this augmentation for Re-ID because the task-related knowledge has not been introduced. Inspired by the research on vulnerabilities of deep neural networks [3, 31], we argue that white box attacks can serve as a bridge between our adjuster and the Re-ID backbone. Therefore, we further utilize it to provide task-related guidance for our data generation.

In addition, we observe that although most augmented samples have natural appearances, many of them inevitably contain artifacts and noise. These subtle flaws bring ambiguity to identity representation learning and cannot be removed by popular domain adaptation methods [1, 44]. To handle this problem, we further introduce **synthetic-to-real feature constraints** for simultaneously narrowing domain gap and improving the robustness of identity features. Experiments on several cross-resolution re-id benchmarks confirm the effectiveness of our approach.

To summarize, our main contributions are as follows:

- We propose a novel global-aware and attack-guided perceptual data generation framework to complement existing training data for Re-ID against the low-level perceptual variations.
- We design a lightweight disentangled generative model which can estimate and manipulate the resolution component of images. It can be easily applied to other perceptual types with few modifications.
- To alleviate the domain bias caused by synthetic samples, we introduce synthetic-to-real feature constraints for narrowing domain gap and regularizing identity feature manifolds.
- The proposed method can be easily integrated into existing deep models without bringing any inference cost. Only in combination with the classic ResNet-50, it can already achieve competitive performance against the state-of-the-art methods on challenging cross-resolution Re-ID benchmarks.

2. Related Work

Conventional person Re-ID. There are various influential visual factors in real-world scenarios, including semantic variations (*e.g*., view, pose, occlusion) and perceptual variations (*e.g*., resolution, illumination), which make person Re-ID a challenging task. Early works exploit local features to alleviate the issues of viewpoints, pose changes and occlusions. These methods adopt attention mechanisms [23, 6, 4], pre-defined regional partition [36, 35, 30] and semantic parts parsing [18, 51] to achieve this goal. However, local features are often unable to deal with low-level perceptual variations that have approximate global uniformity. As a result, Jiao *et al*. [17] firstly combine image super-resolution and Re-ID to solve the resolution mismatch problem. In order to alleviate the difficulty of gradients back-propagation in such a scheme, Cheng *et al*. [9] propose a training regularization strategy, called INTACT, to maximize the compatibility of SR with Re-ID matching. In addition to combining with auxiliary image restoration methods, a series of GAN-based Re-ID methods [7, 24, 16] are proposed to improve identity representation learning. Chen *et al*. [7] attempt to extract resolution-invariant features with adversarial learning. Huang *et al*. [16] propose a degradation invariance learning framework to disentangle the identity contents and low-level visual degradation.

Data generation for Re-ID. In order to alleviate the data shortage issue of the Re-ID task, some works attempt to use generative methods to supplement the training data. Using powerful 3D engines [34], controllable person generation with different poses, backgrounds or illuminations can

be realized, which promotes quantitative analysis of how visual factors influence the Re-ID system. Bak *et al*. [1] employ a game engine to simulate the appearance of hundreds of subjects under different realistic illumination conditions, collecting a new synthetic Re-ID dataset, *i.e*., SyRI. Since GANs has made remarkable progress in face generation [20], a few works [47, 28, 46] attempt to use GANs to augment training data for Re-ID. Zheng *et al*. [47] are the first to introduce GANs into Re-ID for unlabeled sample generation, and a label smoothing regularization is also proposed to leverage unsupervised data. Ma *et al*. [28] propose to generate natural person images by disentangling the input into weakly correlated factors. On the other hand, through intermediate feature interpolation, augmented samples with different poses and backgrounds can also be obtained. For instance, Zheng *et al*. [46] unify generative learning and discriminative learning into a framework. With dynamic soft labeling assignment, the synthesized samples can be used for training in spite of inter-class variations.

3. Methodology

Our approach aims to supplement the training data by adjusting the perceptual degradation of samples, so as to improve the robustness of the Re-ID model against real-world visual degradations. The proposed data augmentation is data-dependent and task-driven, unlike conventional methods [48].

We use two complementary ways to generate auxiliary data, *i.e*., **global-aware augmentation** for densely re-sampling perceptual variations to deal with possible scenarios that are not covered by the training data, **ReID-driven attack** for using task-related knowledge to guide the generation of perceptual adversarial samples which simulate the interference of visual degradation on Re-ID. In order to improve the **identity feature learning** more effectively, these auxiliary data are generated online in each iteration and leveraged by specific constraints that designed for alleviating the sample bias between synthetic domain and real-world domain.

3.1. Network Architecture

As shown in Figure 2, the proposed approach consists of multiple sub-modules, the detailed network structures of which are given in the supplement.

Identity Encoder E_{id} is a ResNet-50 [14] backbone with BNNeck [27] head for identity feature extraction. Only this module takes part in the final inference stage.

Content Encoder E_c is a lightweight convolutional neural network (CNN) with ASPP [5] for extracting content-related image features.

Degradation Encoder E_d is a multi-layer CNN with a normalized linear layer for extracting degradation features and estimating perceptual quality scores of input images.

Degradation Attacker A_d is a stacked graph convolutional network (GCN) [21] with short connections. Taking the degradation features and batch-wise affinity matrix as inputs, the attacker is expected to predict the most intrusive perceptual quality scores.

Generator G is a CNN with adaptive instance normalization (AdaIN) [15] layers which can fuse content and degradation features to form an image.

Discriminator D employs the structure of multi-scale Patch-GAN [50] which is used to distinguish the generated image and encourage the synthetic distribution to be close to the real distribution.

3.2. Global-Aware Perceptual Augmentation

In order to achieve a global-aware data augmentation that obeys to the realistic perceptual distribution of the data source, we attempt to measure the degree of image degradation and estimate the overall distribution. For this purpose, we first need to disentangled the degradation components from images.

Formulation. Given an input image I, corresponding disentangled features f_c and f_d can be obtained by the content encoder and the degradation encoder:

$$f_c = E_c(I), \;\; f_d = E_d(I), \tag{1}$$

while the reconstructed image I_{rec} can be produced by the generator:

$$I_{rec} = G(f_c, f_d). \tag{2}$$

Constraints. As illustrated in Figure 3, to learn such a disentangled generative model, we adopt an image triplet (I^{hr}, I^{lr}, I^{de}) as input, where I^{hr} denotes high-resolution (HR) image, I^{lr} denotes low-resolution (LR) image, while I^{de} is the degraded version of I^{hr} produced by a non-differentiable degraded function, *e.g*., down-sampling for the cross-resolution setting.

By swapping the content features of inputs I^{hr} and I^{de}, the re-generated images can be used to provide disentangled constraints, *i.e*., pixel-wise swap-reconstruction loss:

$$\begin{aligned} L_{rec}^{swap} = &\|G(f_c^{hr}, f_d^{de}) - I^{de}\|_2 \\ &+ \|G(f_c^{de}, f_d^{hr}) - I^{hr}\|_2. \end{aligned} \tag{3}$$

To ensure that the disentangled model is able to reconstruct input images, we use a pixel-wise self-reconstruction loss:

$$\begin{aligned} L_{rec}^{self} = &\|G(f_c^{hr}, f_d^{hr}) - I^{hr}\|_2 \\ &+ \|G(f_c^{lr}, f_d^{lr}) - I^{lr}\|_2 \\ &+ \|G(f_c^{de}, f_d^{de}) - I^{de}\|_2. \end{aligned} \tag{4}$$

Further, we can estimate the perceptual quality score s with the normalized linear layer in degradation encoder:

$$s = L_{norm}(f_d) = f_d \cdot w^T, \tag{5}$$

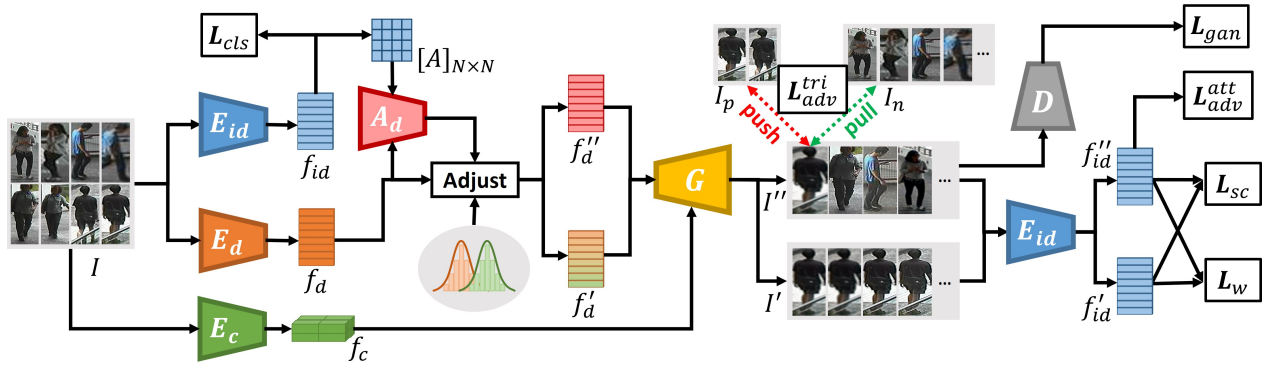


Figure 2. The proposed Global-Aware and Attack-Guided perceptual data generation with synthetic-to-real feature constraints. By adjusting the degradation features of input image I with Eq.(8), global-aware augmentation and ReID-driven attack can be performed to generate I' and I'', respectively. These auxiliary data contribute to identity representation learning and improve the robustness of Re-ID model against low-level perceptual variations.

where w is the linear weights constrained by a score regression loss:

$$L_{reg} = \|s^{hr} - 1\|_2 + \|s^{lr} + 1\|_2. \tag{6}$$

The total objective for the image disentangled generator is formulated as:

$$L_{total}^{Gen} = \lambda_{rec}(L_{rec}^{swap} + L_{rec}^{self}) + \lambda_{reg} L_{reg}, \tag{7}$$

where λ_{rec} and λ_{reg} are balancing weights of losses.

Degradation manipulation. Inspired by the work [32] which found that the latent features of GANs become disentangled and controllable after linear transformations, we use a similar operation to controllably adjust the degradation components of images by:

$$f_d' = f_d + (s' - f_d \cdot w^T) \cdot w, \tag{8}$$

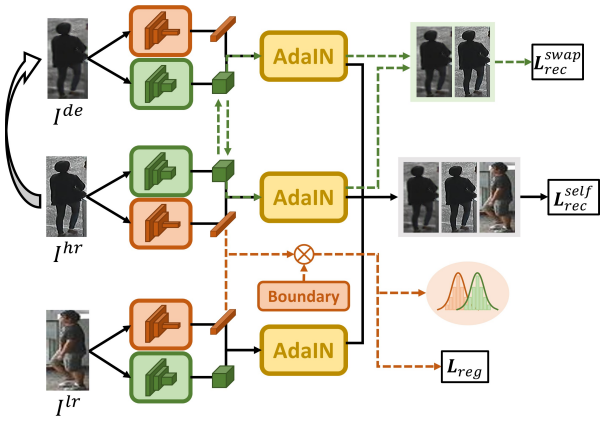


Figure 3. The training process of the proposed lightweight disentangled generative model.

where f_d' is the adjusted degradation feature, s' is the perceptual quality score sampled from the real-world degradation distribution $\mathcal{D}_d$.

Finally, an augmented image I' of the input image I can be produced by:

$$I' = G(E_c(I), f_d'). \tag{9}$$

Discussion. Similar to DI-REID [16], our method relies on content-degradation disentanglement. However, DI-REID aims to learn degradation-invariant identity features explicitly, while our method utilizes global-aware and attack-guided data generation to complement training data, which is more flexible and lightweight.

3.3. ReID-driven Perceptual Attack

Although diverse samples can be obtained by global-aware perceptual augmentation, such a process is independent of high-level tasks and is very inefficient in improving Re-ID performance. We expect that the task-related information can be used as prior knowledge to guide the adjustment of degradation components, so as to simulate the interference of real-world perceptual variations on the Re-ID task. This enlightens us to introduce the white-box attack mechanism with a batch-wise degradation attacker to predict the optimal adjustment values.

Formulation. Specifically, we extract identity feature embeddings e by a well-trained teacher model with the same structure as E_{id}, then an intra-batch affinity matrix $\mathcal{A}$ can be calculated by the normalized euclidean distance metric function:

$$\mathcal{A}_{ij} = 1 - 0.5 \cdot \left\| \frac{e_i}{\|e_i\|} - \frac{e_j}{\|e_j\|} \right\|, \tag{10}$$

where e_i and e_j denotes identity embeddings of the i-th and j-th samples in a mini-batch.

Using $\mathcal{A}$ as the graph adjacency matrix, we then input the degradation features of a mini-batch into the attacker A_d to obtain the ideal perceptual quality score for attacking:

$$s'' = A_d(f_d, \mathcal{A}), \tag{11}$$

which is used to adjust degradation features and generate adversarial samples I'' similar to Eq.(8) and Eq.(9). Note that s'' is scaled by the upper and lower bounds of perceptual score set of the training data, avoiding unreasonable numerical range.

Constraints. In order to simulate the misalignment of identity features and the interference of discrimination cues caused by perceptual variations, we introduce two perceptual attack constraints to optimize the attacker.

To misalign the identity features, a mis-ranking loss function is adopted to minimize the distance of the mismatched pair and maximize the distance of the matched pair:

$$\begin{aligned} L_{adv}^{tri} = \sum_{i=1}^{N} [& \underset{y_k \neq y_i}{max} \|E_{id}(I_i'') - E_{id}(I_k)\| \\ & - \underset{y_j = y_i}{min} \|E_{id}(I_i'') - E_{id}(I_j)\| + \Delta]_+, \end{aligned} \tag{12}$$

where N is the number of samples in a mini-batch, y_i is the identity label of the i-th sample, Δ is the ranking margin. In order to improve the training stability, we calculate the distances between adversarial samples and original samples, instead of the distances within adversarial samples in [38].

As a fine-grained retrieval task, identity recognition in Re-ID relies on local attention, which inspires us to use an attention attack loss:

$$L_{adv}^{att} = \sum_{i=1}^{N} \sum_{l} \left\| \frac{a_l(I_i'')}{\|a_l(I_i'')\|} - \frac{a_l(I_i)}{\|a_l(I_i)\|} \right\|^{-1}, \tag{13}$$

where a_l denotes the attention map of the l-th layer. Following [19], the attention map of layer l is fomulated as:

$$a_l(I) = \sum_{c} |f_{l,c}(I)|^2, \tag{14}$$

where $f_{l,c}$ denotes the c-th channel of the feature maps w.r.t. layer l, and the operations in Eq.(14) are all element-wise. This constraint is expected to misalign attention maps of the adversarial sample and its corresponding original sample.

Besides, a discriminator is adopted to force generated images to be similar to real images:

$$L_{gan} = \mathbb{E}[log(D(I)) + log(1 - D(I''))]. \tag{15}$$

In summary, the total objective for perceptual attack is:

$$L_{total}^{Attack} = \lambda_{adv}^{tri} L_{adv}^{tri} + \lambda_{adv}^{att} L_{adv}^{att} + \lambda_{gan} L_{gan}. \tag{16}$$

3.4. Robust Representation Learning

After acquiring generated samples, we can leverage them to train the Re-ID model together with original samples. However, the domain bias inevitably exists between generated samples and real samples, which makes the learned identity representations deviate from the ideal distribution. To alleviate this issue, we consider the more reasonable synthetic-to-real feature constraints.

Assuming that N_s generated samples $\{I_1^*, I_2^*, ..., I_{n_s}^*\}$ based on the input I have been produced, they can also be encoded into identity embedding $\{e_1^*, e_2^*, ..., e_{N_s}^*\}$. Note that the identity labels of generated samples are the same as that of the input. Let I be an anchor and all I^* as positives, we attempt to keep the anchor fixed and encourage positives to be closer to the anchor. This strategy utilize generated samples to regularize the feature manifold while minimizing their impact on the original identity feature distribution.

Specifically, we introduce two loss functions, *i.e.*, **self-center loss** and **Wasserstein loss**, which force the model to narrow the distances between generated samples and the original sample at the sample level and the instance level, respectively. The self-center loss explicitly encourages the model to push generated samples closer to the original sample in the identity embedding space:

$$L_{sc} = \frac{1}{N} \sum_{i=1}^{N} \max_{j=1,2,...,N_s} ||e(i) - e_j^*(i)||_2^2, \tag{17}$$

where N is the batch size, $e(i)$ is the identity embedding of the i-th sample. The max operation is used as hard sample mining to speed up the training process. Note that the gradient backpropagation of self-center loss to the original embedding e should be detached to alleviate the deviation caused by generated samples.

Despite the simplicity, the self-center loss only focuses on the single-sample case and does not consider the intra-class variations at the instance level. In addition, the calculation cost for self-center loss will increase linearly as N_s becomes larger. To handle these issues, we further introduce a Wasserstein loss to make the distributions of generated samples and original samples as similar as possible for each identity.

Assuming that identity features obey a normal distribution, we are able to perform online estimations [39] to get means and covariance matrices of identity features:

$$\begin{aligned} \tilde{e} &\sim \mathcal{N}(\mu, \Sigma), \\ \tilde{e}^* &\sim \mathcal{N}(\mu^*, \Sigma^*). \end{aligned} \tag{18}$$

The 2-Wasserstein distance is used to measure the similarity of these two Gaussian distributions, which leads to the Wasserstein loss:

$$L_w \triangleq W_2(\tilde{e}, \tilde{e}^*)^2 = ||\mu - \mu^*||_2^2 + ||\Sigma^{\frac{1}{2}} - \Sigma^{*\frac{1}{2}}||_F^2. \tag{19}$$

The corresponding mathematical derivation of L_w is given in the supplement. Note that the loss L_w measures the distance between feature distributions of generated samples and original samples, so it can approximate the ideal situation where $N_s \to +\infty$.

As a result, the total objective for identity representation learning is formulated as:

$$L_{total}^{Id} = \lambda_{cls} L_{cls} + \lambda_{sc} L_{sc} + \lambda_w L_w, \tag{20}$$

where L_{cls} is a standard cross-entropy loss, λ_{cls}, λ_{sc} and λ_w are balancing weights.

4. Experiment

4.1. Datasets

In order to evaluate the Re-ID performance of our approach against low-level perceptual variations, we conduct experiments on four widely used Re-ID datasets, including three cross-resolution benchmarks: MLR-CUHK03 [22], MLR-VIPeR [12] as well as CAVIAR [8] for resolution variations, and a conventional benchmark MSMT17 [41] for multiple variations.

The **MLR-CUHK03** and **MLR-VIPeR** are synthetic cross-resolution datasets based on the person Re-ID benchmarks CUHK03 and VIPeR, respectively. MLR-CUHK03 is composed of 14,097 images from 1,467 identities with 5 different camera views, while MLR-VIPeR includes 632 person image pairs captured by two cameras. Following SING [17], each image from one camera is down-sampled with a ratio randomly picked from $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$ to simulate the cross-resolution situations.

The **CAVIAR** is a genuine cross-resolution dataset which comprises 1,220 images of 72 identities captured by two different cameras in an indoor shopping center in Lisbon. This dataset provides realistic images of multiple resolutions, hence it is very suitable for evaluating real-world cross-resolution person Re-ID.

The **MSMT17** is a very challenging dataset, which covers different dates, time periods, weather conditions, illumination, *etc*. It composes of 32,621/93,820 bounding boxes for training/testing, collected by 15 surveillance cameras on the campus, including outdoor and indoor scenes. Due to the drastic perceptual variations, we use this dataset to evaluate our method against multiple variations.

4.2. Implementation Details

Our approach is implemented in PyTorch with a NVIDIA 1080Ti GPU. All the training and testing images are resized to $256 \times 128 \times 3$, and batch size N is set to 8. For the stage of disentangled generation, the Adam optimizer with learning rate of 0.0001 is used to train the E_c, E_d and G for 50000 iterations. For the stage of perceptual attack, the Adam optimizer with learning rate of 0.01 is adopted to optimize the attacker A_d for 10000 iterations. For the stage of identity representation learning, the SGD algorithm with weight decay of 0.0001 and the Nesterov momentum of 0.9 is used to train E_{id} for 60 epochs. The initial learning rate is 0.02, and it decays to 0.002 after 40 epochs. All balancing loss weights λ_{rec}, λ_{reg}, λ_{adv}^{tri}, λ_{adv}^{att}, λ_{gan}, λ_{cls}, λ_{sc} and λ_w are set to 2.0, 1.0, 2.0, 0.1, 1.0, 1.0, 5.0 and 5.0, respectively. The augmented multiple N_s is set to 4 by default. More details about optimizations and structures can be found in the supplement.

4.3. Re-ID Evaluation and Comparisons

Following the standard evaluation protocols of corresponding datasets, the average Cumulative Match Characteristic (CMC) and the mean Average Precision (mAP) are adopted to evaluate Re-ID performance.

Re-ID against resolution variations. We compare our approach with a wide range of state-of-the-art cross-resolution or conventional Re-ID methods, including (1) SR-based methods: SING [17], CSR-GAN [40], INTACT [9], PRI [13]; (2) Resolution-invariant representation based methods: RAIN [7], DI-REID [16], (3) Hybrid method: CAD [24]. Ohter Re-ID methods, *e.g*., CamStyle [49] and FD-GAN [10], are also compared.

As shown in Table 1, our approach achieves superior performance on all three cross-resolution benchmarks. Specifically, our approach achieves 87.6% at Rank-1 on the MLR-CUHK03 dataset, improves the baseline by 9.5% and outperforms the best competitor INTACT by 1.2%. The identity feature extractor we used is a basic ResNet-50 network without any multi-head or multi-scale designs, hence the performance gain is largely benefited from perceptual data generation and synthetic-to-real feature constraints.

We find that DI-REID [16] achieves the best Rank-1 score on the CAVIAR dataset due to its specified feature disentanglement for the real scenes, while other methods are designed based on the down-sampling operations that are difficult to generalize to the real domain. Even so, our method still achieves the best results on rank-5 and rank-10, and reaches the best accuracy-generalization balance on all three datasets.

Re-ID against the other variations. Although not the main demonstration, the proposed approach shows potential to improve the feature robustness against different types of perceptual variations. Only considering the illumination variations, our approach brings significant performance gains on the challenging MSMT17 dataset, surpassing the baseline model by 9.0% at Rank-1, as shown in Table 2. Since our approach does not depend on the specific network architecture, it is expected to be used in combination with existing state-of-the-art deep models to further improve the performance. We also evaluate our approach on the Market-

Method	MLR-CUHK03			MLR-VIPeR			CAVIAR		
	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
CamStyle [49]	69.1	89.6	93.9	34.4	56.8	66.6	32.1	72.3	85.9
FD-GAN [10]	73.4	93.8	97.9	39.1	62.1	72.5	33.5	71.4	86.5
SING [17]	67.7	90.7	94.7	33.5	57.0	66.5	33.5	72.7	89.0
CSR-GAN [40]	71.3	92.1	97.4	37.2	62.3	71.6	34.7	72.5	87.4
RAIN [7]	78.9	97.3	98.7	42.5	68.3	79.6	42.0	77.3	89.6
CAD [24]	82.1	97.4	98.8	43.1	68.2	77.5	42.8	76.2	91.5
INTACT [9]	86.4	97.4	98.5	46.2	73.1	81.6	44.0	81.8	93.9
DI-REID [16]	85.7	97.1	98.6	50.3	77.9	87.3	**51.2**	83.6	94.4
PRI [13]	85.2	97.5	98.8	-	-	-	43.2	78.5	91.9
Baseline	78.1	93.3	96.0	33.5	59.2	69.9	25.8	62.2	80.0
Baseline + Rand-DS	76.3	93.0	96.5	31.3	58.2	70.3	21.4	57.5	76.5
Ours: Naive†	75.2	92.8	96.4	29.7	46.5	55.7	24.1	59.9	79.7
Ours: w/o Augment	84.8	96.9	97.8	49.7	74.7	84.2	43.8	**84.9**	**95.4**
Ours: w/o Attack	86.1	97.4	98.5	51.3	77.5	84.8	41.6	82.0	92.8
Ours	**87.6**	**97.5**	**99.3**	**52.2**	**79.7**	**88.0**	44.0	84.8	93.6

† Naive means to apply a standard cross-entropy loss to the generated data directly.

Table 1. Cross-resolution Re-ID performance (%) compared to the state-of-the-art methods on the MLR-CUHK03, MLR-VIPeR and CAVIAR benchmarks, respectively. The baseline we used is a ResNet-50 backbone trained with RandomCrop, RandomHorizontalFlip, RandomErasing [48] and BNNeck [27]. Rand-DS denotes randomly down-sample training samples with a probability of 0.5.

Methods	Rank-1	Rank-5	Rank-10	mAP
GoogLeNet [37]	47.6	65.0	71.8	23.0
ResNet-50 [14]	57.4	72.9	78.4	29.2
PDC [33]	58.0	73.6	79.4	29.7
GLAD [42]	61.4	76.8	81.6	34.0
PCB [36]	68.2	81.2	85.5	40.4
Baseline	64.0	78.1	83.0	36.6
Ours-Illumination	**73.3**	**84.9**	**88.5**	**46.6**
Ours-Resolution	71.1	84.0	87.6	44.3
Ours-Blur	71.9	83.8	87.4	45.4

Table 2. Conventional Re-ID performance (%) on the MSMT17 dataset produced by our approach which adopts different degraded function.

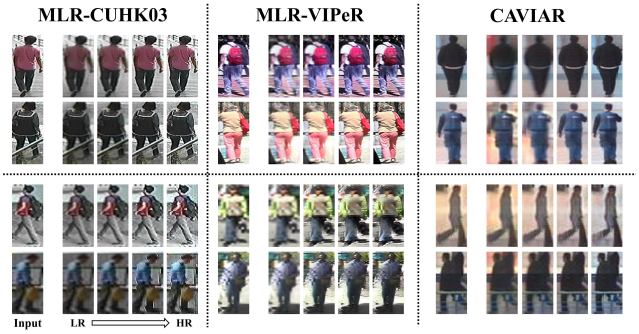


Figure 4. Cross-resolution image augmentation on three Re-ID benchmarks, where all the augmented samples are obtained by linearly interpolating the resolution quality score.

1501 [45] dataset against more types of perceptual variations. These experimental results are presented in the supplement.

Ablation studies. We study the contribution of each component of our approach, as reported in Table 1. It can be observed that directly using our generated data naively with a standard cross-entropy loss cannot improve Re-ID performance. After introducing the self-center loss L_{sc} and the Wasserstein loss L_w, significant performance gains can be achieved, exceeding the naive implementation by 12.4% at Rank-1 on the MLR-CUHK03 dataset. More analysis of losses and hyperparameters are given in the supplement.

4.4. Visualizations

Augmented image generation. To demonstrate the potential of our disentangled generative model, plenty of visualizations of generated samples with resolution and illumination manipulations are given in Figures 4 and 5. By sampling perceptual quality scores linearly, the generated sample sequence has continuous resolution (lighting) changes, which provides rich data diversity for Re-ID training. Different from direct feature interpolation, our method is based on quantitative feature manipulation, which can achieve single-sample augmentation without selecting two samples as endpoints for linear interpolation.

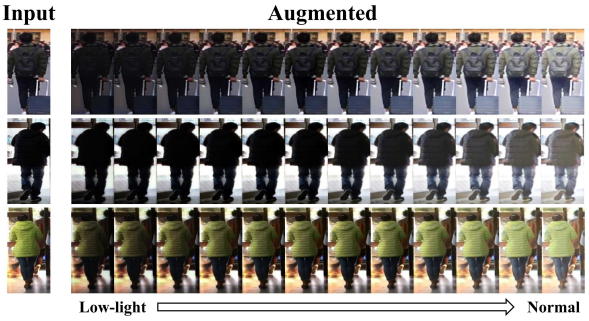
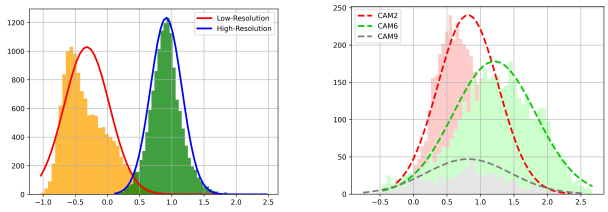


Figure 5. Augmented images against illumination variations.



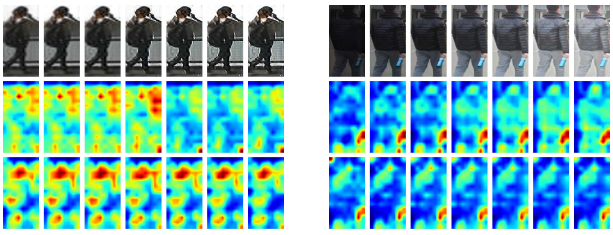
(a) MLR-CUHK03 (b) MSMT17 (illumination)

Figure 6. Perceptual quality distributions. The horizontal axis denotes perceptual quality scores, and the vertical axis denotes number of samples. Solid and dashed lines indicate estimated perceptual score distributions of resolution and illumination, respectively.

Perceptual quality distribution. In fact, as a by-product, our method is able to quantitatively analyze the low-level visual variations present in the Re-ID dataset. As mentioned above, the perceptual quality score of an image can be estimated with Eq(5). Hence, we can calculate the quality scores of all training samples in the dataset and draw the histograms, as shown in in Figure 6. We found that the low-level visual quality distributions of real-world data is very close to the normal distribution, thus it is reasonable to use the Gaussian distribution to approximate the real-world distribution for resampling-based data augmentation.

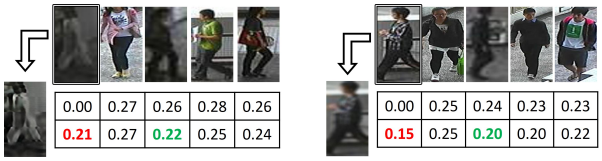
Learned identity features. With our proposed perceptual data generation, the influence of different low-level visual factors on identity features can be quantitatively analyzed, as shown in Figure 7. Compared with the baseline, the identity features extracted by our method show excellent stability under continuously changing resolution and illumination, which further explains why our method can significantly improve the baseline.

Analysis of perceptual attack. To analyze the effect of our perceptual attack, we calculate the normalized Euclidean distance between samples before and after the attack, as shown in Figure 8. It can be observed that after the attack, the distance between the adversarial sample and its original version is significantly increased (marked in Red), while the distance to the negative sample is slightly reduced



(a) Cross-resolution. (b) Cross-illumination.

Figure 7. Visualizations of learned identity features. **Top**: input images, **middle**: features produced by baseline, **bottom**: features produced by our framework.



0.00	0.27	0.26	0.28	0.26
0.21	0.27	0.22	0.25	0.24

0.00	0.25	0.24	0.23	0.23
0.15	0.25	0.20	0.20	0.22

Figure 8. Visualizations of ReID-driven perceptual attack. The values in the table denotes the normalized Euclidean disntances between samples. **First row**: distances before attack, **second row**: distances after attack.

(marked in Green). These hard mismatched pairs generated by attack guidance are expected to promote robust feature learning and improve training efficiency.

5. Conclusion

In this paper, we propose global-aware and attack-guided perceptual data generation to complement existing training data for person Re-ID against low-level perceptual variations. Specifically, we design a lightweight disentangled generative model to estimate and manipulate the perceptual variations of images. We also employ a GCN-based white-box attacker to introduce task-related knowledge for data generation. To alleviate the domain bias caused by synthetic samples, we introduce synthetic-to-real feature constraints to narrow the domain gap and regularize identity feature manifolds. Experiments on four benchmarks demonstrate that our approach effectively improves the Re-ID performance under different low-level perceptual variations.

Acknowledgement

This work was supported by the National Key R&D Program of China under Grand 2020AAA0105702, National Natural Science Foundation of China (NSFC) under Grants U19B2038, the University Synergy Innovation Program of Anhui Province under Grants GXXT-2019-025.

References

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

[2] Igor Barros Barbosa, Marco Cristani, Barbara Caputo, Aleksander Rognhaugen, and Theoharis Theoharis. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *Computer Vision and Image Understanding*, 2018.

[3] Joan Bruna, Christian Szegedy, Ilya Sutskever, Ian Goodfellow, Wojciech Zaremba, Rob Fergus, and Dumitru Erhan. Intriguing properties of neural networks. In *ICLR*, 2014.

[4] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *ICCV*, 2019.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*, 2017.

[6] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *ICCV*, 2019.

[7] Yun-Chun Chen, Yu-Jhe Li, Xiaofei Du, and Yu-Chiang Frank Wang. Learning resolution-invariant deep representations for person re-identification. In *AAAI*, 2019.

[8] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011.

[9] Zhiyi Cheng, Qi Dong, Shaogang Gong, and Xiatian Zhu. Inter-task association critic for cross-resolution person re-identification. In *CVPR*, 2020.

[10] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NIPS*, 2018.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

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

[13] Ke Han, Yan Huang, Zerui Chen, Liang Wang, and Tieniu Tan. Prediction and recovery for adaptive low-resolution person re-identification. In *ECCV*, 2020.

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

[15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[16] Yukun Huang, Zheng-Jun Zha, Xueyang Fu, Richang Hong, and Liang Li. Real-world person re-identification via degradation invariance learning. In *CVPR*, 2020.

[17] Jiening Jiao, Wei-Shi Zheng, Ancong Wu, Xiatian Zhu, and Shaogang Gong. Deep low-resolution person re-identification. In *AAAI*, 2018.

[18] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018.

[19] G. Kang, L. Zheng, Y. Yan, and Y. Yang. Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization. 2018.

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[21] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. 2017.

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

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

[24] Yu-Jhe Li, Yun-Chun Chen, Yen-Yu Lin, Xiaofei Du, and Yu-Chiang Frank Wang. Recover and identify: A generative dual model for cross-resolution person re-identification. In *ICCV*, 2019.

[25] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. Adaptive transfer network for cross-domain person re-identification. In *CVPR*, 2019.

[26] Jiawei Liu, Zheng-Jun Zha, Xuejin Chen, Zilei Wang, and Yongdong Zhang. Dense 3d-convolutional neural network for person re-identification in videos. *ACM TOMM*, 2019.

[27] 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.

[28] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018.

[29] Dechao Meng, Liang Li, Xuejing Liu, Yadong Li, Shijie Yang, Zheng-Jun Zha, Xingyu Gao, Shuhui Wang, and Qingming Huang. Parsing-based view-aware embedding network for vehicle re-identification. In *CVPR*, 2020.

[30] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, 2019.

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

[32] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020.

[33] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.

[34] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*, 2019.

[35] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *CVPR*, 2019.

[36] 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.

[37] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[38] 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 *CVPR*, 2020.

[39] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. In *NIPS*, 2019.

[40] Zheng Wang, Mang Ye, Fan Yang, Xiang Bai, and Shin'ichi Satoh. Cascaded sr-gan for scale-adaptive low resolution person re-identification. In *IJCAI*, 2018.

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

[42] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *ACMMM*, 2017.

[43] Wei Zhang, Shengnan Hu, Kan Liu, and Zhengjun Zha. Learning compact appearance representation for video-based person re-identification. *IEEE TCSVT*, 2018.

[44] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *CVPR*, 2018.

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

[46] 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.

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

[48] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020.

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

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[51] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *ECCV*, 2020.