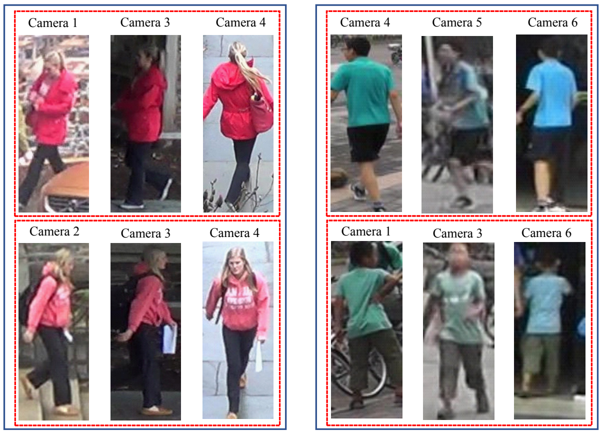
Camera-Aware Image-to-Image Translation Using Similarity Preserving StarGAN For Person Re-identification

Dahjung Chung Edward J. Delp
Video and Image Processing Laboratory (VIPER),
School of Electrical and Computer Engineering,
Purdue University, West Lafayette, Indiana, USA
chung123@purdue.edu ace@ecn.purdue.edu

Abstract

Person re-identification is a crucial task in intelligent video surveillance systems. It can be defined as recognizing the same person from images of a person taken from different cameras at different times. In this paper, we present a camera-aware image-to-image translation using similarity preserving StarGAN (SP-StarGAN) as the data augmentation for person re-identification. We propose the addition of an identity mapping term and a multi-scale structural similarity term as additional losses for the generator. SP-StarGAN can learn the relationship among the multiple cameras with a single model and generate the camera-aware extra training samples for person re-identification. We evaluate our proposed method on public datasets (Market-1501 and DukeMTMC-reID) and demonstrate the efficacy of our method. We also report competitive performance with the state-of-the-art methods.



(a) DukeMTMC-reID [40] (b) Market-1501 [38]

Figure 1: Sample Images showing challenges related to camera variations in the ReID problem

1. Introduction

The number of video surveillance system has grown exponentially in recent years making the continues monitoring of surveillance data impossible [1]. To automate the analysis of the surveillance data, intelligent video surveillance system has been an active research area in computer vision. The goal is to extract meaningful information efficiently from the large volume of surveillance data.

Person re-identification (ReID) is one of the fundamental task associated with intelligent video surveillance system. ReID refers to tracking a person across a network of non-overlapping cameras [5, 12]. Given single/multiple images or a video sequence of the interested person (query), ReID is the task of recognizing the same person within the list of images/videos collected from multiple cameras with non-overlapping field of view (gallery).

Even though ReID has been intensively studied over the past years, it is still an active research area due to various challenges. ReID dataset can have intensive illumination changes, pose variations, occlusions, different scales and camera viewpoints [12]. In addition, collecting and annotating a large dataset for multiple cameras is a very time-consuming and expensive process.

Figure 1 shows the challenges in ReID. It demonstrates that it is challenging to distinguish the same person in the images taken from different cameras. Most of these challenges are due to camera variations such as different settings or environments of multiple cameras. In order to address these challenges, many ReID methods have been proposed by adopting new features, using metric learning techniques, the use of semantic attributes and using deep learning approaches.

Recent traditional ReID approaches have focused on appearance modeling and metric learning to learn the representations that can be invariant to cameras-related properties such as illumination and view point changes [21,25,26]. There have been traditional approaches proposed including KISSME [21], XQDA [25] and GOG [26]. More recently, deep learning based methods have been described such as IDE [39], Two Stream Siamese Net [8], SVDNet [29], Re-Ranking [41] and TJ-AIDL [33].

As deep learning approaches have been studied, large-scale ReID datasets have been released : Market-1501 [38] and DukeMTMC-reID [40]. Compare to the other datasets such as PRID [17], iLIDS-VID [34], Market-1501 and DukeMTMC-reID have more than 6 different cameras settings and a large number of images and identities. This means that the same person can show up in more than two camera views which introduces more challenges to re-identify the person.

Although larger datasets have been introduced, more training data is needed. In addition, since the number of camera is growing in these datasets, more samples for each cameras are needed in order to learn robust camera invariant feature representations. It is expensive and time-consuming to have manual identity annotations across different cameras as we have more cameras and videos to annotate the identity for ReID tasks.

To alleviate this problem, Zhong *et al.* [43] proposed a method for generating camera style-transferred images using a CycleGAN (CamStyle) [44] as a data augmentation method for ReID. They trained multiple image-to-image translation models for each camera pair using CycleGAN. Then, the model can generate new sample images from the source camera style to the target camera style. Camera style means the camera specific settings such as bright or dark illumination. These style-transferred images allow us to have extra training samples with different camera styles without additional manual annotation. In addition, label smoothing regularization (LSR) on the style-transferred images to softly distribute their labels and reduce the noise effect generated by the extra samples. Due to the limitation of CycleGAN which can model only one-to-one domain mapping, this method only can learn the mapping for one camera pair (e.g., camera 1 to camera 2) with the single model. Thus, using Camstlye [44], multiple models need to be trained to model an entire camera network. For example, in the DukeMTMC-reID [40] dataset which has 8 different cameras, $C_8^2 = 28$ different models need to be trained separately. The time complexity as well as the the number of parameters will be sharply increased as the number of camera increases. In addition, cross-camera relationships will be ignored in this architecture. To address these limitations, we propose the use of StarGAN [7] with an additional similarity preserving term in the loss function for camera-aware image-to-image translation to generate the extra samples for ReID.

The main contributions of this paper are :

- We propose a similarity preserving StarGAN (SP-StarGAN) which is an improvement of StarGAN [7]. To improve the quality of the generated images, we propose to add a identity mapping term and multi-scale structural similarity (MS-SSIM) to the generator loss. SP-StarGAN can be used not only for ReID data augmentation but also for general multi-domain image-to-image translation. Compare to the previous method (Camstyle), our method has almost 15 times less parameters to train while producing competitive generated image quality as well as competitive accuracy in ReID.

- For ReID, we use SP-StarGAN to generate more training samples across different camera settings. In addition, we propose to employ the Re-Ranking method [41] for ReID as post processing along with SP-StarGAN generated samples in order to improve ReID matching accuracy. We demonstrate that Re-Ranking shows higher performance in ReID accuracy with better quality generated images.

2. Related work

Generative Adversarial Networks. Goodfellow *et al.* [13] proposed the Generative Adversarial Networks (GANs) which learns generative models through an adversarial process that is training a generative model and a discriminative model simultaneously. In recent years, GANs has been used many applications including image generation [27] and image-to-image translation [7, 19, 44]. Radford *et al.* [27] introduced deep convolutional generative adversarial networks (DCGANs) that have some architectural constraint for stable training and they have demonstrated the applicability for image generation. One of the extensions of GANs, Pix2Pix [19], used a conditional GANs to learn the relationship between the output and input image for image-to-image translation. Pix2Pix [19] has the limitation that it requires paired training data. To overcome this limitation, a coupled generative adversarial network (CoGAN) was introduced to learn the joint distribution across domains without having the paired training data. Next, cycle consistency adversarial networks (CycleGAN) employed a cycle-consistency term in the adversarial loss for image-to-image translation without having paired samples. CycleGAN has the limited scalability that it can only learn the mapping between two domain. This requires multiple models to be trained in order to translate images across multiple domains. To address this problem, Choi *et al.* [7] proposed a unified generative adversarial networks (StarGAN) which allows us to learn the mapping between multiple domains with a single model.

Deep learning Approaches in ReID. With the release of larger datasets, [23] demonstrated the feasibility of using of

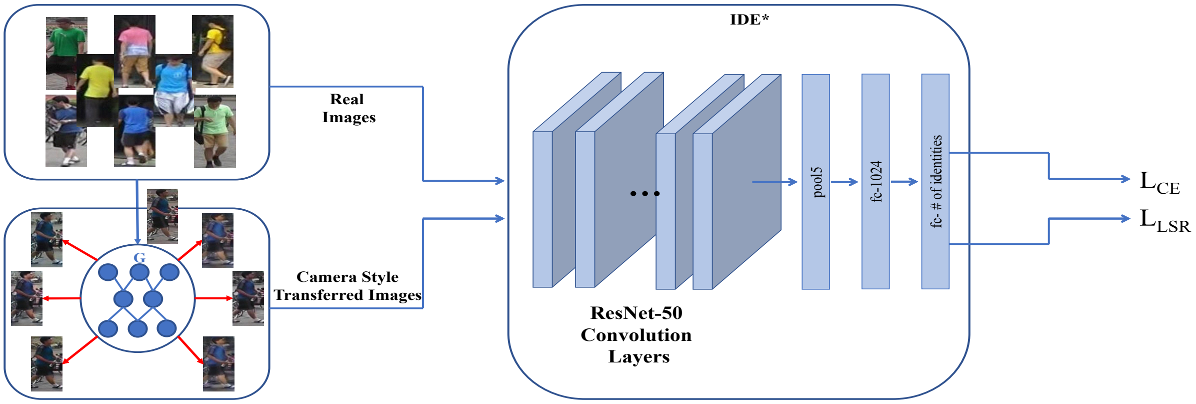


Figure 2: Overall Proposed Framework

deep learning for ReID by using a filter pairing neural network (FPNN). After this initial work, many deep learning methods were proposed to improve ReID accuracy [2,6,37]. Ahmed *et al*. [2] proposed to use a cross-input neighborhood difference layer to compute the differences in feature values across different camera views. In [37], a cosine layer connects two sub-networks and jointly learn color, texture and a similarity metric. Later, Cheng *et al*. [6] employed a multi channel CNN to jointly learn both global and local features of the human body using triplet loss function. In [32], a Siamese Long Short-Term Memory (LSTM) model that can process image parts sequentially is described. The use of LSTM enables the capability of memorizing the spatial dependency and selectively propagating the context information throughout the network. Chung *et al*. [8] proposed the use of a two stream Siamese network to learn the spatial and temporal feature representation for ReID. In addition, Zheng *et al*. [39] proposed ID-Discriminative Embedding (IDE) using ResNet [15] to train a ReID model as a classifier. In [41], they propose a re-ranking method using k-reciprocal Encoding inspired by [4]. This method can be used with any initial ranking.

More recently, Deng *et al*. [10] proposed image-to-image translation across different dataset domains while preserving self similarity and domain dissimilarity. Their method consists of an Siamese network and a CycleGAN. In [33], Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) was introduced to learn attribute-semantic and identity discriminative features simultaneously and transfer them to the target domain without collecting additional data from the target domain. In order to reduce the dataset domain discrepancy, Wei *et al*. [36] proposed a Person Transfer Generative Adversarial Network (PTGAN) using a CycleGAN to learn the relationship between two different dataset domain.

Even though larger ReID datasets are available, the number of samples are still limited to train CNN models due to the expensive annotation process. Thus, over-fitting still can happen due to the lack of training samples in ReID dataset. To address this problem, some data augmentation methods have been proposed [40,42,43]. Zhong *et al*. [42] proposed a random erasing technique which randomly selects the rectangle region and erases it with random values to avoid the over-fitting problem. In [40], DCGAN [27] was used to generate unlabeled person images. They also proposed the label smoothing regularization for outliers (LSRO) to assign the stable labels for the generated images. More recently, Zhong *et al*. [43] introduced a camera style transferred image generation using CycleGAN [44] as a data augmentation method for ReID. They also described improved label smoothing regularization (LSR) for generated images to address small portion of unreliable data. We will refer this method [43] as Camstyle for the rest of the paper.

3. Proposed Method

Figure 2 shows the overall flow of our proposed method. First, we train the similarity preserving StarGAN to obtain the camera-aware image-to-image translation model. This model learns the mapping across different cameras with a single model in the ReID dataset. We then generate camera style translated images for all respective camera combinations from this single model. Finally, we train the deep learning ReID network with both real images and camera style translated images.

3.1. StarGAN

In this section we briefly revisit the StarGAN [7]. StarGAN has a single generator G learning the mappings among multiple domains and a single discriminator D with auxiliary classifier to discriminate fake and real images and control multiple domain simultaneously. In order to stabilize the training process while generating realistic fake images, the Wassertein GAN loss with a gradient penalty

[3, 14] was used for the adversarial loss and defined as:

$$
\begin{aligned}
L_{adv} = &\mathbb{E}_x[D_s(x)] - \mathbb{E}_{x,c_t}[D_s(G(x, c_t))] \\
&- \lambda_{gp}\mathbb{E}_{\hat{x}}[(||\nabla_{\hat{x}} D_s(\hat{x})||_2 - 1)^2]
\end{aligned}
\tag{1}
$$

where D_s is defined as the probability distribution over the sources, $\hat{x}$ is uniformly sampled along a straight line between a pair of a real and a generated image. In addition, G generates an image $G(x, c_t)$ mapped from the input image x to the target domain label c_t, while D tries to distinguish the between real and generated images.

StarGAN [7] has an auxiliary classifier on top of D to classify images to the respective domain label. For the real image, a domain classification loss is defined as:

$$
L^r_{cls} = \mathbb{E}_{x,c_s}[-logD_{cls}(c_s|x)] \tag{2}
$$

where $D_{cls}(c_s|x)$ denotes the probability distribution over domain labels given the real image x and c_s means the source domain label. For the fake image, a domain classification loss is described as:

$$
L^f_{cls} = \mathbb{E}_{x,c_t}[-logD_{cls}(c_t|G(x, c_t))] \tag{3}
$$

where $D_{cls}(c_t|G(x, c_t))$ represents a probability distribution over domain labels given the fake image $G(x, c_t)$ and c_t refers the target domain label.

In order to preserve the content of the input images while translating the domain-related information of the image, StarGAN used a cycle consistency loss [44] which is defined as

$$
L_{rec} = \mathbb{E}_{x,c_t,c_s}[||x - G(G(x, c_t), c_s)||_1] \tag{4}
$$

where the translated image $G(x, c_t)$ becomes the input for the G with the original domain label c_s and reconstruct the original image x.

Finally, the overall StarGAN loss function is expressed as

$$
L_D = -L_{adv} + \lambda_{cls}L^r_{cls}, \tag{5}
$$

$$
L_G = L_{adv} + \lambda_{cls}L^f_{cls} + \lambda_{rec}L_{rec} \tag{6}
$$

where λ_{cls}, λ_{rec} are hyper-parameters for the relative importance of each term.

3.2. Similarity Preserving StarGAN

In this paper, we employ the StarGAN model to generate camera style translated images as extra training samples for ReID. However, we observe that the cycle consistency term, L_{rec} in Equation 4, was not enough for preserving the content of the input image related to person identity while translating the camera domain-related content. For camera-aware image-to-image translation, we do not want to have dramatic changes in the image since we need to keep the same identity while transferring the different camera settings. In order to preserve the same identity while transferring the image to the different camera setting, we add two additional terms into StarGAN generator loss (Equation 6). We present the details of each additional term in the following.

Identity Mapping Loss. In order to preserve the color-consistency between the input and output, we add the identity mapping loss [30] to regularize the generator to be an identity mapping when the real image with the source domain label is provided.

The identity mapping loss term is defined as

$$
L_{id} = \mathbb{E}_{x,c_s}[||G(x, c_s) - x||_1] \tag{7}
$$

where $G(x, c_s)$ is the generated image with the source camera label c_s and the x is the real image from camera c_s.

Multi-scale Structural Similarity. Wang *et al.* [35] originally used the structural similarity between two images across different scales. We add the MS-SSIM term to our generator loss to preserve the structural similarity. Specifically in camera-aware image translation, we need to preserve the most of the structural information to maintain the same identity. By using this term, the generator tries to preserve the structural information of the input image.

Let $x_r = G(G(x, c_t), c_s)$ as the reconstructed image with the source camera label c_s, c_t refers to the target camera label and the x as the input image. The SSIM loss can defined as

$$
L_{SSIM}(x_r, x) = [l(x_r, x)^{\alpha}c(x_r, x)^{\beta}s(x_r, x)^{\gamma}] \tag{8}
$$

where

$$
l(x_r, x) = \frac{2\mu_{x_r}\mu_x + C_1}{{\mu_{x_r}}^2 + {\mu_x}^2 + C_1} \tag{9}
$$

$$
c(x_r, x) = \frac{2\sigma_{x_r}\sigma_x + C_2}{{\sigma_{x_r}}^2 + {\sigma_x}^2 + C_2} \tag{10}
$$

$$
s(x_r, x) = \frac{\sigma_{x_r x} + C_3}{\sigma_{x_r}\sigma_x + C_3}. \tag{11}
$$

$l(x_r, x)$, $c(x_r, x)$, $s(x_r, x)$, α, β, γ represent the luminance, contrast and structure information and their relative importance, respectively. μ_{x_r}, μ_x are the means of x_r and x and σ_{x_r}, σ_x are the standard deviations of x_r and x. $\sigma_{x_r x}$ is the covariance of x_r and x and $C_1 = 0.01^2$, $C_2 = 0.03^2$, $C_3 = C_2/2$ are the fixed hyper-parameters.

We defined MS-SSIM [35] as

$$
\begin{aligned}
L_{MS-SSIM}(x_r, x) = &[l_M(x_r, x)]^{\alpha_M} \\
&* \prod_{i=1}^{M}[c(x_r, x)]^{\beta_i}[s(x_r, x)]^{\gamma_i}.
\end{aligned}
\tag{12}
$$

Full SP-StarGAN loss function. Finally, the proposed full generator loss function to optimize can be defined as

$$L_G = L_{adv} + \lambda_{cls} L^{f}_{cls} + \lambda_{cls} L_{rec} \\ + \lambda_{id} L_{id} - \lambda_s L_{MS-SSIM} \qquad (13)$$

where λ_{cls}, λ_{rec}, λ_{id}, λ_s are the relative importance of domain classification, reconstruction, identity mapping and MS-SSIM losses, respectively. Note that we have the same discriminator loss function as in Equation 5.

Network Architecture. We employ the same network architecture from [7]. The generator is consist of two convolutional layers with stride size 2, 6 residual blocks [15] and two transposed convolutional layers with stride size 2. The instance normalization [31] was used only for the generator. For the discriminator, the PatchGANs [22] was used to classify the local image patches are real or fake.

3.3. Deep Person ReID Network

Base Deep ReID Model. We use the ID-Discriminative Embedding (IDE) [39] to train ReID model. In this network, we use ResNet-50 [15] convolutional layers followed by global max pooling layer. We then add two fully connected layers as stated in [43]. The first layer has 1024 dimensions followed by batch normalization [18], ReLU and Dropout [28]. For the ID-Discriminative Embedding, we have the second layer that has P (the number of class dimensions) in order to use cross-entropy loss.

Loss Function. We use the cross-entropy loss for the real images. For the generated images, we utilize the label smoothing regularization (LSR) as suggested in [43] to reduce the negative effect of some of the noisy generated images. Even though we have the identity label for the generated images, some images have transfer noise due to the occlusions or the noise in the input image. To alleviate this problem, LSR assigns the small weights to the other classes and give less confidence in the identity label.

Re-Rank. We employ the re-ranking method [41] as post processing on our initial ranking results from base deep ReID model. Zhong *et al.* proposed to use the k-reciprocal encoding for ReID re-ranking. Re-rank computes features by encoding its k-reciprocal neighbors into a single vector. Then this vector is used to re-rank under the Jaccard distance. And the final distance is computed with the combination of the original distance and the Jaccard distance. We will refer this method as Re-Rank in the rest of the paper.

4. Experiments

4.1. Datasets

Market-1501 [38] contains 32,668 images in total with 1,501 identities from 6 different camera views. From the video, person images were detected using a deformable part model [11]. This dataset is partitioned into 12,935 images (751 identities) for training and 19,732 images (750 identities) for the gallery. In ReID test, 3,3668 hand-captured images from 750 identities are pre-selected as queries to evaluate ReID performance. Single-query evaluation protocol is used.

Method	λ_{id}	λ_s	mAP	Top-1 Rank
Baseline (IDE*)	-	-	65.87	85.66
StarGAN	0	0	66.1	86.5
StarGAN + Identity	1	0	67.2	86.7
	2	**0**	**68.2**	**87.9**
	5	0	67.4	87.5
StarGAN + MS-SSISM	0	1	67.4	87.4
	0	**2**	**67.6**	**87.4**
	0	5	66.2	85.9
StarGAN + Both	1	1	67	87.2
	1	2	67.6	87.5
	1	5	67.6	86.9
	2	**1**	**68.6**	**88.1**
	2	2	68.5	87.6
	2	5	67.6	87.6

Table 1: ReID accuracy evaluation on different proposed components in SP-StarGAN loss on Market-1501

DukeMTMC-reID [42] has 36,411 images in total with 1,404 identities from 8 different camera views. It is composed of 16,522 images (702 identities) for training samples and 17,661 images (702 identities) for the gallery. In ReID test, 2,228 images from 702 identities are pre-selected as queries for the evaluation.

4.2. Experiment Setup

Similarity Preserving StarGAN. We first resize the images to 178x178 and then crop them randomly to 128x128. The horizontal random flip is used with a probability of 0.5 as the data augmentation. As described in [7], all models are trained using Adam [20] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The generator updates after five discriminator updates as in [3]. The initial learning rate is 0.0001 for the first 100,000 iterations and linearly decays to the learning to 0 over the next 100,000 iterations. The batch size is 16. We use the fixed hyper-parameter values for $\lambda gp = 10$, $\lambda cls = 1$, $\lambda rec = 10$ in Equation 13. We describe the analysis to select the best value for λ_{id}, λ_s in Section 4.3.

Finally, for inference, we generate all combinations of different cameras per image with the image size 128x128. For example, if the real image is taken from camera 1 and we have K different cameras in the dataset, then generate the translated images with target camera domain label from 2 to K.

Component	Base Augmentation		Base + RE		Base + Re-Rank		Base + RE + Re-Rank	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
StarGAN	66.1	86.5	69.8	88.4	82.1	88.9	85.6	90.9
StarGAN + Identity	68.2	87.9	71	88.9	83.3	89.8	86.7	91.3
StarGAN + MS-SSIM	67.6	87.4	71	88.7	82.8	89.3	85.6	91.2
StarGAN + Both	68.6	88.1	70.9	88.5	83	89.5	86.3	91.1

Table 2: ReID accuracy evaluation on different proposed pre/post processing methods on Market-1501

Deep Person ReID Network. We follow the training general strategy in [43] to train the base deep reID model except for the learning rate policy. All images are resized to 256x128. Two base data augmentation method were used for the training : random cropping and random horizontal flipping. A model is trained with SGD solver. The initial learning rate is set to 0.01 for the ResNet-50 convolutional layers and 0.1 for the two additional fully connected layer since we use ImageNet [9] pre-trained ResNet-50 layers as the initialization. In our experiments, the initial learning is divided by 10 after first 30 epochs out of 60 epochs in total. The batch size is set to 128 and the dropout probability is set to 0.5.

In the ReID test, we extract the feature from the pooling layer and use Euclidean distance to compute the similarity between the gallery and query images. We use the generated images as the extra training samples and follow the strategy of [43] in the selection of the generated images. We randomly select M real images and N generated images in a training mini-batch. We set the $M : N$ ratio to $3 : 1$ for all experiments. We evaluate the ReID performance in terms of mean Average Precision (mAP) and Top-1 Rank matching accuracy.

4.3. Component Evaluation

In this section, we investigate the significance of the components in GAN part and ReID Network Part of the proposed method.

Similarity Preserving StarGAN. We first investigate the effect of the additional loss terms in SP-StarGAN on ReID accuracy metrics. We evaluate for different hyper-parameter settings such as StarGAN + identity, StarGAN + MS-SSIM and StarGAN + Both when λ_{id} and λ_s are varying from $1 - 5$. Note that this evaluation was done without any additional augmentation or post-processing in Deep Re-ID network. In [43], they defined IDE* with the improved learning rate policy while keeping the same network architecture from IDE [39]. IDE* is used as the baseline to evaluate the proposed components. As shown in Table 1, the usage of original StarGAN [7] improved around 1% from the baseline in both mAP and Top-1 Rank accuracy. When we included the additional loss terms into the generator loss function, we obtain around 2% improvement in ReID accuracy depending on the hyper-parameters λ_{id} and λ_s. This improvement is coming from the generating better quality images which results that having less noise in generated samples. We also observe that we do not have the continuing improvement as we increase the contributions of the additional loss terms.

Deep Person ReID Network. We evaluate the different components in Deep ReID network including Random Erasing (RE) and Re-Rank. For this evaluation, we fix the hyper-parameters for the GAN part as $\lambda_{id} = 2$ and $\lambda_s = 1$. For any type of proposed GANs, we observe the significant improvement in ReID accuracy by employing both RE and Re-Rank. This result demonstrates that using Random Erasing as extra data augmentation along with the Re-Rank as the post processing has significant positive effect on ReID accuracy. Thus, our final proposed method version in the following section will be including both RE and Re-Rank as well as StarGAN + Both method with the parameters $\lambda_{id} = 2$ and $\lambda_s = 1$.

4.4. Complexity Analysis

Table 3 shows the comparison of the complexity of the model between CamStyle [43] and proposed method. Note that this experiment was done using a NVIDIA Titan Xp GPU. CamStyle has around 792 M parameters to train while our proposed method has only 52.23 M parameters to train as shown in Table 3a. For training and inference processing time as shown in Table 3b, CamStyle takes around 150 more hours in training than the proposed method for DukeMTMC-reID [40] dataset. Camstyle can only learn the mapping between two different camera domains at one time due to the limitation of CycleGAN. This results the dramatic increase in the complexity since we need to train multiple models. On the other hand, proposed method can model the mapping between multiple camera domains with the single model while showing the competitive ReID accuracy.

4.5. Comparisons

Visual Evaluation Comparison We compare the sample generated images from Camstlyle and our proposed

Sub-Network	Number of Parameters [M]	
	CamStyle	Ours
Generator	637.17 M	8.44 M
Discriminator	154.84 M	44.79 M
Total	792.01 M	53.23 M

(a) Number of Parameters on DukeMTMC-reID [40]

Mode	Processing Time [hours]	
	CamStyle	Ours
Training	304.17	12.84
Inference	2.16	0.12

(b) Processing Time on DukeMTMC-reID [40]

Table 3: A complexity comparison on CamStyle [43] and Our Proposed Method

method. Both Camstlye and our proposed method can generate competitive quality of person images. However, as shown in Figure 3, in this particular sample, proposed method can generate better quality images especially in person's leg compare to Camstyle [43] and the original StarGAN [7]. This particular sample has a lot of noise in the input image and it demonstrates that proposed method can create better quality image even with the noisy input.

ReID Evaluation Comparison For the full version of proposed method, we use the StarGAN + Both where $\lambda_{id} = 2$ and $\lambda_s = 1$ with RE and Re-Rank. We compare our proposed method with the state-of-the-art methods on Market-1501 and DukeMTMC-reID in Table 4 and 5. In both datasets, our proposed method outperforms all the other methods in terms of both mAP and Top-1 Rank accuracy. We achieve significant improvement in especially mAP (15-72%) by employing Re-Rank with SP-StarGAN. We also achieve the highest accuracy in terms of Top-1 Rank accuracy in both datasets.

5. Conclusions

In this paper, we propose the camera-aware multiple domain image-to-image translation using Similarity Preserving StarGAN (SP-StarGAN) for person re-identification(ReID). We propose the SP-StarGAN which has identity mapping loss and Multi-scale Structural Similarity loss in the generator loss function. The SP-StarGAN can learn the mapping among all different camera settings in ReID dataset and generate the camera-aware translated images as the extra training samples in ReID with a single model. We demonstrate that having two additional loss terms helps address the quality problem in generated images as well as ReID performance. Our experimental results also demonstrate that by using SP-StarGAN along with Random Erasing and Re-Rank improves the ReID performance. In the future, we want to extend this work to cross-domain ReID problem.

Methods	mAP	Top-1 Rank
LOMO + XQDA [25]	14.09	34.4
IDE [39]	46	72.54
Re-rank [41]	63.63	77.11
SVDNet [29]	62.1	82.3
TriNet [16]	69.14	84.92
DJL [24]	65.5	85.1
DCGAN [40]	66.07	83.97
IDE* [43]	65.87	85.66
IDE* + CamStyle [43]	68.72	88.12
IDE* + CamStyle + RE [42]	71.55	89.49
Ours (full version)	86.3	91.1

Table 4: A ReID accuracy comparison on Market-1501

Methods	mAP	Top-1 Rank
BOW + KISSME [21]	12.17	25.13
LOMO + XQDA [25]	17.04	30.75
IDE [39]	44.99	65.22
SVDNet [29]	56.8	76.7
TriNet [16]	72.44	53.5
DCGAN [40]	47.13	67.68
IDE * [43]	51.83	72.31
IDE * + CamStyle [43]	53.48	75.27
IDE* + CamStyle + RE [42]	57.61	78.32
Ours (full version)	65	82.1

Table 5: A ReID accuracy comparison on DukeMTMC-reID

References

[1] Cisco visual networking index: Forecast and methodology, 2015/2020. *Cisco Systems Inc.*, April 2016.

[2] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3908–3916, June 2015. Boston, MA.

[3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. *Proceedings of the International Conference on Machine Learning*, 70:214–223, Aug 2017. Sydney, Australia.

[4] S. Bai and X. Bai. Sparse contextual activation for efficient visual re-ranking. *IEEE Transactions on Image Processing*, 2016.

[5] A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286, April 2014.

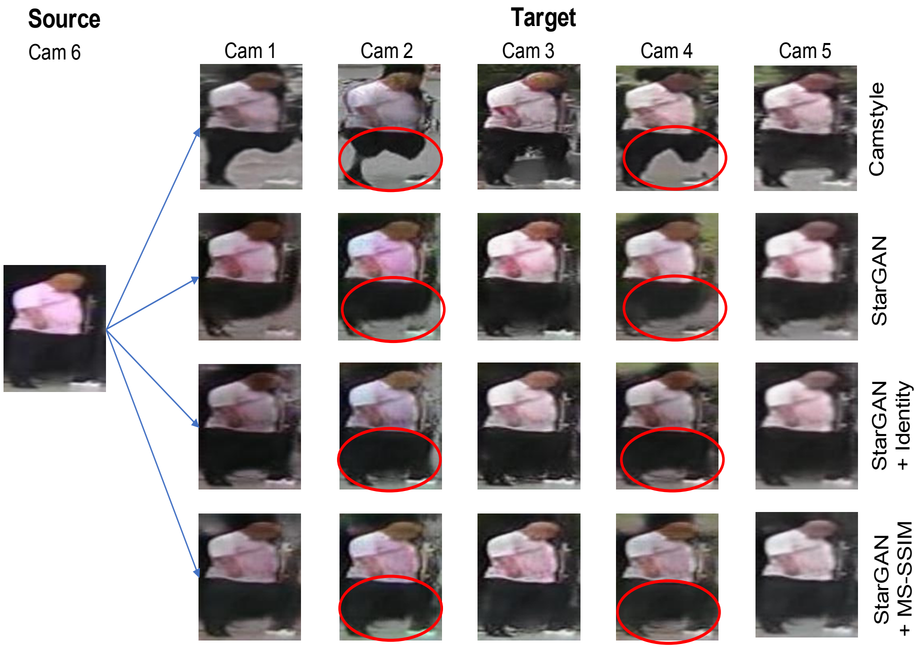


Figure 3: Sample Generated Image Comparison

[6] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, June 2016. Las Vegas, NV.

[7] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[8] D. Chung, K. Tahboub, and E. Delp. A two stream siamese convolutional neural network for person re-identification. *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017. Venice, Italy.

[9] J. Deng, W. Dong, R. Socher, L. Li, and and. Imagenet: A large-scale hierarchical image database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[10] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. Salt Lake City, UT.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sep 2010.

[12] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person re-identification*. Springer, London, 2014.

[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Proceedings of the International Conference on Neural Information Processing Systems*, pages 2672–2680, 2014. Montreal, Canada.

[14] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans. *Proceedings of the International Conference on Neural Information Processing Systems*, pages 5769–5779, 2017. Long Beach, CA.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. Las Vegas, NV.

[16] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *ArXiv preprints*, 2017.

[17] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. *Proceedings of the Scandinavian Conference on Image Analysis*, pages 91–102, May 2011. Ystad, Sweden.

[18] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the International Conference on Machine Learning*, pages 448–456, Jul 2015. Lille, France.

[19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision*, July 2017. Honolulu, Hawaii.

[20] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *Proceedings of the International Conference on Learning Representations*, 12 2014.

[21] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2288–2295, June 2012. Providence, RI.

[22] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. *Proceedings of the European Conference on Computer Vision*, Oct 2016. Amsterdam, Netherlands.

[23] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, June 2014. Columbus, OH.

[24] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. pages 2194–2200, 2017. Melbourne, Australia.

[25] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, June 2015. Boston, MA.

[26] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1363–1372, June 2016. Las Vegas, NV.

[27] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *Proceedings of the International Conference on Learning Representations*, abs/1511.06434, 2016. Vancouver, Canada.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[29] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017. Venice, Italy.

[30] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. *Proceedings of the International Conference on Learning Representations*, April 2017. Toulon, France.

[31] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *ArXiv preprints*, 2016.

[32] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. Las Vegas, NV.

[33] J. Wang, X. Zhu, S. Gong, and W. Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. Salt Lake City, UT.

[34] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. *Proceedings of the European Conference on Computer Vision*, pages 688–703, September 2014. Zurich, Switzerland.

[35] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, 2:1398–1402, Nov 2003.

[36] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. Salt Lake City, UT.

[37] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. *Proceedings of the International Conference on Pattern Recognition*, pages 34–39, August 2014. Stockholm,Sweden.

[38] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. *Proceedings of the IEEE International Conference on C omputer Vision*, 2015.

[39] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *ArXiv preprints*, 2016.

[40] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *Proceedings of the IEEE International Conference on Computer Vision*, July 2017. Honolulu, Hawaii.

[41] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. *Proceedings of the IEEE International Conference on Computer Vision*, July 2017. Honolulu, Hawaii.

[42] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *ArXiv preprints*, 2017.

[43] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. Salt Lake City, UT.

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017. Venice, Italy.