Semantic Consistency and Identity Mapping Multi-Component Generative Adversarial Network for Person Re-Identification

Amena Khatun Simon Denman Sridha Sridharan Clinton Fookes
Image and Video Laboratory, Queensland University of Technology (QUT), Brisbane, QLD, Australia
Email: {a2.khatun, s.denman, s.sridharan, c.fookes}@qut.edu.au

Abstract

In a real world environment, person re-identification (Re-ID) is a challenging task due to variations in lighting conditions, viewing angles, pose and occlusions. Despite recent performance gains, current person Re-ID algorithms still suffer heavily when encountering these variations. To address this problem, we propose a semantic consistency and identity mapping multi-component generative adversarial network (SC-IMGAN) which provides style adaptation from one to many domains. To ensure that transformed images are as realistic as possible, we propose novel identity mapping and semantic consistency losses to maintain identity across the diverse domains. For the Re-ID task, we propose a joint verification-identification quartet network which is trained with generated and real images, followed by an effective quartet loss for verification. Our proposed method outperforms state-of-the-art techniques on six challenging person Re-ID datasets: CUHK01, CUHK03, VIPeR, PRID2011, iLIDS and Market-1501.

1. Introduction

Person re-identification (Re-ID) aims to match an image of a person to a large gallery set, where probe and gallery images are from different cameras. Although person Re-ID is a widely investigated research area, it is still challenging to re-identify the target person accurately in the presence of domain variations including changes due to illumination, pose, viewing angle, and background. Thus in a real world scenario where the domain of the target images has no overlap with the gallery images, performance is severely reduced. To address the domain variation challenge, previous researchers adopted feature extraction methods to learn discriminative features across different cameras. However, while these methods have helped relax the closed-world assumptions of previous methods, performance is degraded when confronted with a real-world scenario where target image conditions are unseen.

Motivated by this problem, we propose a multi-component generative adversarial network for style adaptation, from one to many domains, to improve the discriminative ability of a CNN trained for person Re-ID. Specifically, for image domain translation, a multi-component model is proposed to generate synthetic images where the style of a person is transferred from one domain to multiple other domains with an identity mapping loss and a semantic consistency loss. The identity mapping loss is used to ensure that the identity of the transferred person is the same as the original person, and the semantic consistency loss is used to encourage the network to preserve the learned embedding during domain translation. Thus the Re-ID model will be trained with a larger set consisting of real and synthetic images for the same person showing different styles, such as changes in background and lighting.

Recently, CycleGAN [53] has been used by person Re-ID researchers to transfer style from camera-to-camera or domain-to-domain, however CycleGAN can only transfer images between two domains at once. In contrast, our pro-

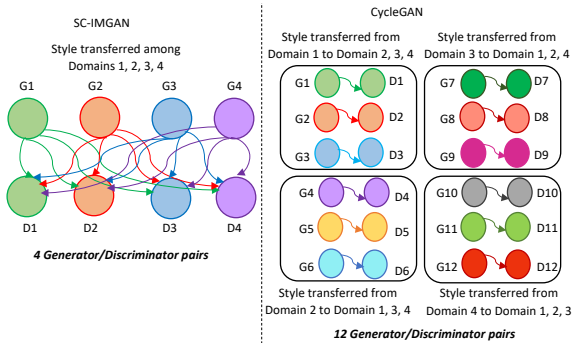

Figure 1: Illustration of the proposed SC-IMGAN and the existing CycleGAN [53]. CycleGAN translates images between two domains at a time; while SC-IMGAN can transfer images between multiple domains with an identity mapping loss to ensure that images retain their identity after translation, and a semantic consistency loss to preserve semantic information shared across domains. Thus for four domains, CycleGAN requires 12 generator/discriminator pairs while SC-IMGAN requires only 4 pairs.

posed SC-IMGAN is able to generate new images with the style of multiple domains at a time, all with the same identity as the original image. As shown in Figure 1, let us assume we have 4 domains. To transfer styles among these 4 domains, CycleGAN requires 12 generator/discriminator pairs whereas the proposed SC-IMGAN requires only 4 generator/discriminator pairs. Our work also differs in that the generator of SC-IMGAN aims to keep the same identity after translation between domains. CycleGAN uses only the cycle-consistency loss between the real and the reconstructed images at the pixel-level, and as such fails to capture semantic features shared across domains. This leads to a drop in performance when domains are vastly different. To address this limitation, we propose a semantic-consistency loss operating at the feature-level (i.e. on the embeddings learnt by the encoder) to ensure consistent semantic information is extracted for both the input and style transferred images. These newly generated unlabeled synthetic samples are then used as additional positive images to train the Re-ID network alongside real images. Hence, the trained network benefits as it learns different appearance variations (pose, lighting and background) for a person.

Within the proposed framework, we adopt the verification-identification approach of [18], with a quartet loss. However, the quartet loss of [18] does not specify how close the intra-class features should be in feature space, resulting in a drop in performance due to large intra-class distances between images of the same identity. As such, we propose an improved quartet loss which forces the network to minimise the intra-class distance more than the inter-class distance, regardless of whether the positive and negative pairs share the same probe image or not, and simultaneously ensures intra-class features are close to each other, improving network generalisation. Without this, images of the same class may form a large cluster with a relatively large average intra-class distance in feature space which is not desirable. The contributions of this paper are:

- We propose to generate synthetic images with a multi-component generative adversarial network where the images from one domain are transferred to all other available domains simultaneously.
- We demonstrate how identity can be better preserved during style transfer by using an identity mapping loss and a novel semantic consistency loss.
- We propose a novel improved quartet loss to minimise the distance between images of the same identity more than the distance between dissimilar identities, improving the generalisation of the Re-ID network.
- We exceed the state-of-the-art accuracy compared to existing methods on six challenging person Re-ID datasets: CUHK01 [21], CUHK03 [22], VIPeR [15], PRID2011 [16], iLIDS [34] and Maerket-1501 [46].

2. Related Work

In this section, we briefly summarise related research in image domain translation using GANs, and deep learning based person Re-ID.

2.1. Image Domain Translation by Generative Adversarial Networks (GANs)

Generating realistic synthetic images from real images is a challenging task, and requires a model to capture the distribution of the real images. To generate synthetic images which have similar properties to the original training dataset, GANs were introduced in [14]. Generally, GANs use noise to synthesise an image and the network is trained in an adversarial manner. Inspired by the success of GANs, various extensions have been proposed for image-to-image translation [13, 28], pixel-level transfer from source to target domains [43], and style transfer between domains [42]. Rather than using noise alone as the stimulus, the conditional GAN (cGAN) [17] is proposed to control the mode of generated images, however, cGANs need a pair of images for training which is not available for many tasks. To address this, [26] introduced coupled GANs which use a pair of GANs instead of a pair of images. CycleGAN [53] also overcomes the requirement of paired data through a cycle consistency model, where the source domain image is translated according to the target domain and vice-versa. However, CycleGAN can transfer the styles only between two domains at a time. Similar to CycleGAN, [42] proposed DualGAN for unpaired image-to-image translation using dual learning to train the translator network with two sets of images from two domains. Although in StarGAN [9], a multi-domain translation network is proposed using a single generator which takes one-hot vector along with each input to represent domain information, this method is only applied when there is no feature mismatch between domains such as face attribute modification, where all the domains have slight shifts in qualities of the same category of images: human faces with a clear background. Moreover, the restrictive nature of modeling all mapping function as a single network may create problems when the mapping functions between different pairs of domains varies.

As the performance of person Re-ID drops severely due to variations between domains or cameras, researchers adopted GANs for image translation to generate synthetic images with different styles so that CNNs can be trained with multiple styles of a person. Re-ID researchers [48, 50, 36] have typically adopted the traditional GAN [14] or CycleGAN [53] to generate synthetic images which are used to train a Re-ID network. In [48], the GAN is first introduced for person Re-ID to generate new samples which are not present in the training data. However, [48] only generated new samples for data augmentation instead of increasing the number of positive pairs. To translate the images between two domains, [50, 36, 51] employed CycleGAN, aiming to

find a mapping function between two domains. Although these methods achieve promising performance, they don't consider moving beyond two domains. As most camera networks contain 10's or even 100's of cameras, the ability to transfer between an arbitrary number of domains is required. As preserving person identity is crucial for Re-ID, [36, 3] propose adding an identity preserving loss using a foreground mask, however, they require an additional network and extra supervision to extract the mask images.

2.2. Deep Learning For Person Re-ID

As deep CNN (DCNNs) combine feature extraction and metric learning [2, 52] in a single framework, person Re-ID researchers adopted DCNNs to achieve state-of-the-art performance. Siamese networks are adopted by [31, 32, 22], and a pair of images are taken as input and the network is trained to push images of the same identity close to each other in feature space. Other researchers adopted a triplet network which minimises the intra-class distance with respect to the same probe image. [8] improved the original triplet loss by adding new constraints to further minimise intra-class distance using a second margin; and [6] further improved the triplet loss through the use of two margins: the first performing the same function as the original triplet loss, while the second seeks to maximise inter-class distance. However, the second margin is weaker than the first, which leads to the network being dominated by the triplet loss, i.e. minimising the intra-class distance when the probe images belong to the same person. To address this problem, a new loss is proposed in [18] to force the network to minimise the intra-class distance more than the inter-class distance, regardless of whether the probe image comes from the same person or not. However, [18] does not specify how close the positive pair should be in feature space. In contrast to [6, 18], we propose a new loss function for Re-ID which not only reduces the intra-class distance more than the inter-class distance with respect to multiple different probe images, but also specifies how close the positive pair should be in feature space.

A number of recent approaches have used part based methods [20, 45, 4, 40] with the aim of overcoming occlusions, and handling partial observations. These methods, however, all rely on verification or identification frameworks only instead of jointly adopting both approaches, and some methods require additional supervision and pose annotation. Other methods [12, 44, 23] are focused on clustering or transfering the knowledge from a labeled source dataset to an unlabelled data using pseudo labels. However, different identities may have the same pseudo label which can make it hard for the model to distinguish similar people.

Our work differs from the above approaches in architecture, loss function and motivation. Rather than only address the issue of image translation between two domains, we propose a multi-component adversarial network to transfer the style from one to many domains at once. To improve person Re-ID performance, we add an identity mapping loss to preserve the identity of transferred images. In addition to the cycle consistency loss applied at pixel level, we propose a semantic-consistency loss applied at the feature-level to capture shared semantic content for flexible cross-domain image translation. The real and style transferred images are then fed into the proposed four-stream CNN with the improved quartet loss for verification, and softmax loss for identification. Finally, as in a real world situation gallery and query images likely have no overlap, we build a cross-domain architecture to cope with such a scenario.

3. Proposed Method

The proposed architecture is shown in Figure 2, and consists of two networks: one for image domain translation; and one for Re-ID. These networks are explained in the following subsections.

3.1. Semantic Consistency and Identity Mapping Multi-Component GAN

The widely adopted CycleGAN learns to map between only two domains at a time. We propose an identity mapping and semantic feature preserving multi-component adversarial network to address the problem of mapping images when more than two domains exist. Let us assume that we have N source domains: $S^1, S^2, S^3, ...S^N$. Thus the proposed method learns to find a mapping among all available domains. For N domains, the proposed method requires N generators and discriminators where each of the generators contains an encoder and decoder. To compare the distribution of the generated images to the distribution of other domains, adversarial losses are used. For example, if we want to transfer the style of domain $S^1 \rightarrow S^2$, the adversarial loss is given by:

$$L_{GAN}(E_1, G_1, D_2, S^1, S^2) = {}_{x_2 \sim P_{S^2}}\,[logD_2(x_2)] + {}_{x_1 \sim P_{S^1}}[log(1 - D_2(G_1(E_1(x_1))))], \quad (1)$$

where the mapping function is $G_1(E_1) : S^1 \rightarrow S^2$ and D_2 is the discriminator. The generator attempts to generate images with the distribution of the new domain (S^2) and the discriminator D_2 tries to differentiate generated images from real images. However adversarial training requires paired training data, otherwise infinitely many mappings will induce the same distribution over the output; and thus many input images will map to the same output image in the absence of paired training data. To address this problem, we adopt the cycle consistency loss [53] to translate images from domain S^1 to domain S^2, and then translate it back from domain S^2 to domain S^1, and as such do not require paired training data. For example, two domains require two mapping functions which should be bijective. The cycle consistency loss can be expressed as,

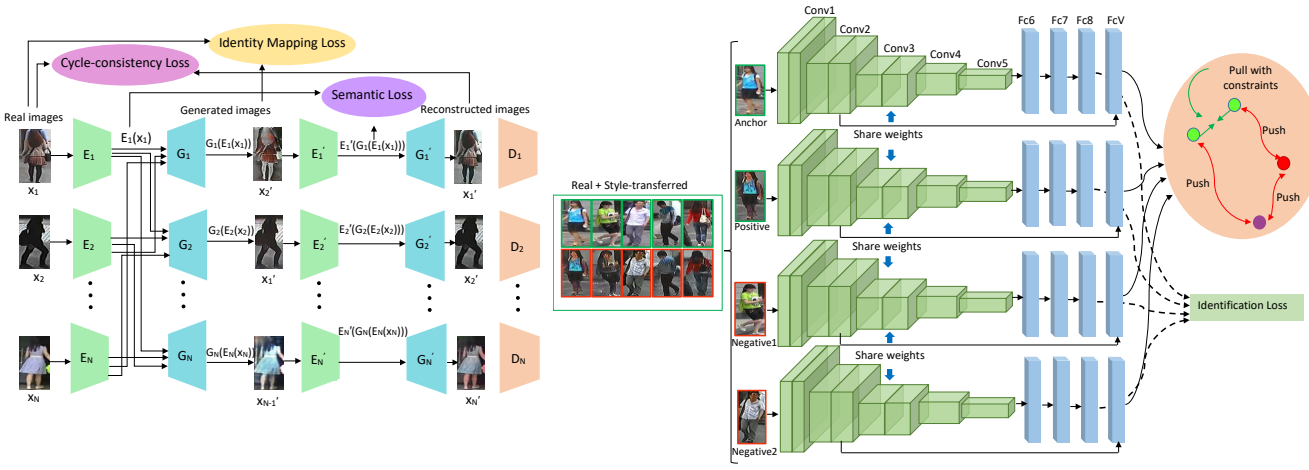

Figure 2: Architecture of the proposed model. At first, style transferred images are generated by SC-IMGAN. The cycle-consistency loss and identity mapping loss are applied at the pixel level while the semantic-consistency loss is appiled at the feature level. We use six domains (CUHK01, CUHK03, VIPeR, PRID2011, iLIDS and Market-1501) to transfer the styles of pedestrians. The style transferred images are then concatenated with real images to train the proposed four-stream deep CNN model. The extracted features are fed into the verification and identification losses to identify the person.

$$L_{cyc} = {}_{x_1 \sim P_{S^1}} [||G_1'(E_1'(G_1(E_1(x_1)))) - x_1||_1] + {}_{x_2 \sim P_{S^2}} [||G_2'(E_2'(G_2(E_2(x_2)))) - x_2||_1], \quad (2)$$

where $G_1'(E_1'(G_1(E_1(x_1))))$ represents the reconstructed version of the real image x_1, and $G_2'(E_2'(G_2(E_2(x_2))))$ is the reconstructed version of the real image, x_2 .

As preserving person identity is crucial for person Re-ID, we add an identity mapping loss [30] alongside the cycle consistency loss to force the generator to preserve the identity of the source domain's real images in the target domain, i.e. we require that if the source domain shows an image of person p, then the person p is also rendered in the target domain. The identity preserving loss can be expressed as,

$$L_{identity} = {}_{x_1 \sim P_{S^1}} [||G_1(E_1(x_1)) - x_1||_1] + {}_{x_2 \sim P_{S^2}} [||G_2(E_2(x_2)) - x_2||_1], \quad (3)$$

where $G_1(E_1(x_1))$ and $G_2(E_2(x_2))$ are the style transferred images from the real images, x_1 and x_2 respectively.

Further, we propose a feature-level semantic-consistency loss to preserve semantics during cross-domain translation, helping maintain identity between vastly different domains. The semantic-consistency loss is given by,

$$L_{semantic} = {}_{x_1 \sim P_{S^1}} [||E_1'(G_1(E_1(x_1))) - E_1(x_1)||_1] + {}_{x_2 \sim P_{S^2}} [||E_2'(G_2(E_2(x_2))) - E_2(x_2)||_1]. \quad (4)$$

This loss is applied to the embeddings so that the encoder extracts the same high level features for both the input and output, such that semantic information is consistent across domains. Here, $E_1'(G_1(E_1(x_1)))$ represents the embedding of the translated images from domain S_1 to S_2 and $E_1(x_1)$ is the embedding of the S_1 domain's real images, and similar for the reverse mapping. Thus to transfer the style from domain S^1 to S^2 with the preserved identity and semantic features, the objective of SC-IMGAN is,

$$L_{SC-IMGAN} = L_{GAN}(E_1, G_1, D_2, S^1, S^2) + L_{GAN}(E_2, G_2, D_1, S^2, S^1) + \lambda_1 L_{cyc} + \lambda_2 L_{identity} + \lambda_3 L_{semantic}. \quad (5)$$

When training the proposed multi-component network, we train the mapping between two domains at a time, and iterate through pairs of domains to train all mappings, with the aim of preserving semantic information and person identity. In this work, we consider six domains (CUHK01, CUHK03, VIPeR, PRID2011, iLIDS and Market-1501) for image translation which requires 6 generator/discriminator pairs. Each of the generators are disengaged to utilize half as encoders and the other half as decoders. Thus for each domain, an encoder and a decoder from different domains can be combined to reduce the required number of generators. As such, the network can translate an image from one domain to all other domains.

3.2. Joint Verification-Identification for Person Re-Identification

The newly generated style transferred images are used as the input alongside real images in a CNN for person Re-ID. Here, a four-stream DCNN is proposed to combine verification and identification tasks in a single framework. We propose an improved quartet loss for verification which requires four input images denoted as, $I_i = I_i^1, I_i^2, I_i^3, I_i^4$ where I_i^1 is the anchor image, I_i^2 is the positive image, and I_i^3 and I_i^4 are two different negative images. Although great success has been had with the triplet loss for person Re-ID,

it suffers from poor generalisation in real world scenarios due to totally unseen target data. The triplet loss pushes images of the same identity close to each other only when the probe images come from the same identity, which is not practical in the real world. We also notice that neither the triplet or quartet loss specify how close the positive pair should be in feature space. Thus intra-class variation within feature space may be high, resulting in a severe drop in performance. To overcome these challenges, we propose an improved quartet loss to minimise the intra-class variation over the inter-class variation, regardless of whether the probe image belongs to the same person or not. We further insert a new term to push the network to minimise the intra-class distance more than the inter-class distance, and ensure this distance is less than a second margin. The full objective of the proposed verification loss is given by,

$$L_{ImpQuartet} = \sum_{i=1}^{n} \Big(max\{ \parallel \Theta_w(I_i^1) - \Theta_w(I_i^2) \parallel^2 - \parallel \Theta_w(I_i^1) - \Theta_w(I_i^3) \parallel^2 + \parallel \Theta_w(I_i^1) - \Theta_w(I_i^2) \parallel^2 - \parallel \Theta_w(I_i^4) - \Theta_w(I_i^3) \parallel^2, \tau_1 \} + max\{ \parallel \Theta_w(I_i^1) - \Theta_w(I_i^2) \parallel^2, \tau_2 \} \Big). \quad (6)$$

Here, the positive pair, comprised of $\Theta_w(I_i^1)$ and $\Theta_w(I_i^2)$, is included twice in the first term of Equation 6 to compensate for having two negative pairs ($\Theta_w(I_i^1)$, $\Theta_w(I_i^3)$; and $\Theta_w(I_i^4)$, $\Theta_w(I_i^3)$). The first negative pair shares a common probe image with the positive pair (i.e. $\Theta_w(I_i^1)$), while the second negative pair uses two different images. Thus, the proposed loss forces the network to maximize the inter-class distance even if the target image comes from a different identity, while the second term forces the distance between $\Theta_w(I_i^1)$ and $\Theta_w(I_i^2)$ to be less than τ_2, where τ_2 is less than τ_1, ensuring that features for the same identity are close in feature space. Thus by using the improved quartet loss, the inter-class distance is required to be larger than the intra-class distance irrespective of whether the probe image comes from the same identity or multiple different identities; while ensuring that the intra-class features will lie close to each other in the feature space.

The identification model in this work is the same architecture as [18, 7] which forces the network to push images of different identities away from each other to identify the query image. As the proposed model trains the network with four input images, three pairs are created for identification, one is a positive pair and the remaining two are negative pairs. For identification, a softmax layer is used to obtain the similarity between the probe and the gallery images. The identification loss is,

$$L_{identification} = -\sum_{i=1}^{n} p_i log\bar{p}_i = -log\bar{p}_t, \quad (7)$$

where p_i is the probability distribution of the target, i is the number of classes, $\bar{p}_i$ is the predicted probability distribution and t is the target class.

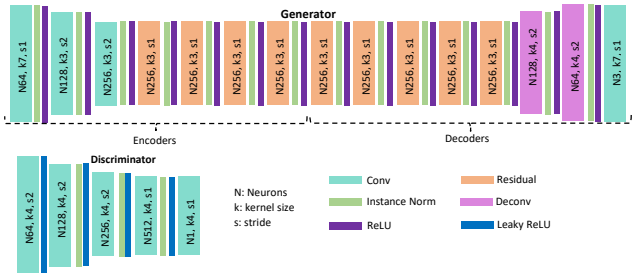

Figure 3: Architecture of the GAN where N, k, and s represents the number of neurons, kernel size and stride respectively. For six domains, our method requires six generator/discriminator pairs, i.e. an encoder, a decoder and a discriminator per domain.

According to [33], the lower layers of deep architectures encode more discriminative features to capture intra-class variations and provide more detailed local features and the higher layers of deep architectures focus on identifiable local semantic concepts. We borrowed this idea and compare the images based on low level features for verification and extract identification features from higher layers as in our work, we need to focus on intra-class variations for verification purpose and local semantic concepts for identification.

3.3. Training the Network

For style-transferred image generation, we use the training data from six person Re-ID databases to train SC-IMGAN. For each training image, we map to each target domain, such that the network learns how to map each image to each domain. We use pytorch and we empirically set λ_1 o 10, and λ_2 and λ_3 to 0.1 (see Equation 5). The Adam optimizer is used to train SC-IMGAN from scratch with a batch size of 1 and a learning rate of 0.0002. We train for 200 epochs. The learning rate is constant for the first 100 epochs, and then linearly decays towards zero over the next 100 epochs. The styles are transferred among 6 domains which requires 6 generator/discriminator pairs whereas CycleGAN requires 30 generator/discriminator pairs, i.e. 30 training networks. To train for 200 epochs, the proposed SC-IMGAN takes approximately 120 hours whereas CycleGAN takes around 1560 hours to train the generators, on a single GPU. The generator and discriminator architecture of SC-IMGAN is illustrated in Figure 3. The encoder contains three convolutional layers. The output activation is then passed through a series of nine residual blocks, which is expanded by the decoder.

For person Re-ID, we use a four stream CNN, fine-tuned from an Alexnet [19] model pre-trained on ImageNet [10]. We set the learning rate to 0.001 and use a batch size of 128 during training. τ_1 is set to -1 and τ_2 to 0.01 in Equation 6.

Method	Type	CUHK01			CUHK03			VIPeR			PRID2011			iLIDS			Market1501		
		R1	R5	R10	R1	R5	R10	R1	R5	R10	R1	R5	R10	R1	R5	R10	R1	R5	R10
DML [41]	ID	-	-	-	-	-	-	28.2	59.2	73.4	17.9	37.5	45.9	-	-	-	-	-	-
FPNN [22]	ID	27.9	-	-	20.7	51.3	68.7	-	-	-	-	-	-	-	-	-	-	-	-
DRDC [11]	V	-	-	-	-	-	-	40.5	60.8	70.4	-	-	-	52.1	68.2	75.1	-	-	-
FLCA [27]	ID	46.8	71.8	80.5	-	-	-	42.5	72	91.7	-	-	-	-	-	-	-	-	
DGD [38]	ID	71.7	88.6	92.6	75.3	-	-	38.6	-	-	64	-	-	64.6	-	-	-	-	-
Improved Trp [8]	V	53.7	84.3	91.0	-	-	-	47.8	74.7	84.8	22.0	47.0	57.0	60.4	82.7	90.7	-	-	-
SIRCIR [32]	ID+V	71.8	-	-	52.2	-	-	35.8	-	-	-	-	-	-	-	-	-	-	-
DRPR [5]	V	70.9	92.3	96.9	-	-	-	38.3	69.2	81.3	-	-	-	-	-	-	-	-	-
PersonNet [37]	V	71.1	90	95	64.8	89.4	94.9	-	-	-	-	-	-	-	-	-	-	-	-
DLCNN [47]	ID+V	-	-	-	83.4	97.1	98.7	-	-	-	-	-	-	-	-	-	-	-	-
GSCNN [31]	ID	-	-	-	68.1	88.1	94.6	37.8	66.9	77.4	-	-	-	-	-	-	65.9	-	-
CAN [24]	ID	-	-	-	72.3	93.8	98.4	-	-	-	-	-	-	-	-	-	60.3	-	-
Re-ranking [49]	ID	-	-	-	61.6	-	-	-	-	-	-	-	-	-	-	-	77.1	-	-
DCF [20]	ID	-	-	-	74.2	94.3	97.6	-	-	-	-	-	-	-	-	-	80.3	-	-
SSM [1]	V	-	-	-	76.6	94.6	98.0	53.7	-	91.5	-	-	-	-	-	-	82.2	-	-
EDM [29]	ID	69.4	-	-	61.3	-	-	40.9	-	-	-	-	-	-	-	-	-	-	-
MTDNet [7]	ID+V	77.5	95.0	97.5	74.7	95.9	97.5	45.9	71.9	83.2	32.0	51.0	62.0	-	-	-	-	-	
BTL [6]	V	62.6	83.4	89.7	75.5	95.1	99.1	49.0	73.1	81.9	-	-	-	-	-	-	-	-	-
P2S [52]	V	77.3	93.5	96.7	-	-	-	-	-	-	70.7	95.1	98.9	-	-	-	70.7	-	-
Spindle Net [45]	ID	79.9	94.4	97.1	88.5	97.8	98.6	53.8	74.1	83.2	67.0	89.0	89.0	66.3	86.6	91.8	76.9	91.5	94.6
DaRe [35]	V	-	-	-	73.8	-	-	-	-	-	-	-	-	-	-	-	90.9	-	-
AACN [39]	ID	88.1	96.7	98.2	91.4	98.9	99.5	-	-	-	-	-	-	-	-	-	88.7	-	-
MLFN [4]	ID	-	-	-	82.8	-	-	-	-	-	-	-	-	-	-	-	90.0	-	-
DFSN [18]	ID+V	83.9	98.2	98.9	85.5	98.7	99.8	68.7	88.9	94.6	75.0	93.0	97.0	-	-	-	-	-	-
CamStyle [50]	ID	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	89.5	-	-
SC-IMGAN + Imp Quartet	ID+V	**95.3**	**99.5**	**99.8**	**92.8**	**99.8**	**99.8**	**72.8**	**95.1**	**98.0**	**80.7**	**96.2**	**99.4**	**86.0**	**96.2**	**98.3**	**94.7**	**95.7**	**98.5**

Table 1: Comparison of SC-IMGAN with state-of-the-art methods on CUHK01, CUHK03, VIPeR, PRID2011, iLIDS and Market-1501 datasets. ID, V and ID+V indicate that an identification, a verification or a combination is used. R1, R5 and R10 indicates rank-1, rank-5 and rank-10 identification accuracy respectively.

The training iterations are set to 30,000 and training takes approximately 24 hours. Stochastic gradient descent (SGD) is used to update network parameters.

4. Experimental Results and Discussions

4.1. Datasets

For SC-IMGAN, six Re-ID datasets (CUHK01, CUHK03, VIPeR, PRID2011, iLIDS, and Market1501) are used to train the generator network where the training images (training splits are the same as discussed below) are used to generate synthetic images. For each dataset, the generated synthetic images are used to train the Re-ID network alongside real images.

For Re-ID, we use the above mentioned six datasets separately to evaluate the proposed method; i.e. for the CUHK01 dataset, the Re-ID network is trained with the real images from CUHK01 dataset along-with the generated images from CUHK01 in the style of other domains. CUHK01 [21] consists of 3884 images of 971 persons taken by different two camera views, each identity has four images. CUHK03 [22] consists of 1467 identities, captured by six surveillance cameras from a university campus. VIPeR [15] contains 1264 images of 632 identities, captured by different cameras with changes in viewing angles, poses and lighting conditions. PRID2011 [16] contains 385 and 749 identities in two camera views which are captured from videos. Among 1134 persons, only 200 are common to both camera views. iLIDS [34], consists of 479 images of 119 identities which are extracted from video images in a busy airport environment. Market-1501 [46] consists of 12936 training images of 751 identities and 19732 testing images of 750 identities which is close to a real world setting. For all six datasets, we follow the same settings for training and testing as [45].

4.2. Comparison with state-of-the-art approaches

We compare the results of our proposed method with state-of-the-art approaches as shown in Table 1. For CUHK01, the proposed SC-IMGAN with the improved quartet loss achieves 95.3% rank-1 accuracy whereas the previous state-of-the-art approach achieved 88.1%, which used an identification task only. For CUHK03 and Market-1501, the proposed method achieves 92.8% and 94.7% rank-1 accuracy respectively. The previous best method achieved 91.4% on CUHK03, though they used additional semantic information such as pose estimation for identification and rely on body part detectors; and 90.9% on Market-1501 where features from multiple layers are combined to capture both high-resolution and semantic details and adopted the traditional triplet loss. For VIPeR, PRID2011 and iLIDS, we outperform the previous state-of-the-art methods by 4.1%, 5.7% and 19.7% rank-1 accuracy respectively.

4.3. Cross-domain Evaluation

For person Re-ID, cross domain experiments are more relevant for real world deployments. To evaluate whether the domain gap is reduced by SC-IMGAN, we trained the network on CUHK03 and Market1501 and tested on PRID2011 as summarised in Table 2 . From Table 2, the model is trained with images from CUHK03 transferred to other domains by SC-IMGAN alongside real images, and achieves significant performance gains when tested on PRID, e.g., a 13.5% and 10.5% increase in rank-1 accuracy compared to [36]. Similar improvements can be observed when trained with the Martket1501 dataset transfered to other domains, outperforming state-of-the-art approaches

Method	CUHK03 → PRID				Market1501 → PRID			
	cam1/cam2		cam2/cam1		cam1/cam2		cam2/cam1	
	R1	R10	R1	R10	R1	R10	R1	R10
PTGAN(cam1) [36]	18.0	43.5	6.5	24.0	17.5	50.5	8.5	28.5
PTGAN(cam2) [36]	17.5	53.0	22.5	54.0	10.0	31.5	10.5	37.5
ATNet(cam1) [25]	-	-	-	-	24.0	51.5	21.5	46.5
ATNet(cam2) [25]	-	-	-	-	15.0	51.0	14.0	41.5
SC-IMGAN (cam1)	**31.5**	**50.0**	**20.0**	**32.5**	**28.0**	**56.5**	**26.0**	**49.5**
SC-IMGAN (cam2)	**28.0**	**57.5**	**36.5**	**60.5**	**21.5**	**55.0**	**20.5**	**45.5**

Table 2: Cross-domain performance comparison on PRID2011 dataset trained with CUHK03 and Market1501 dataset. $cam1/cam2$ indicates that $cam1$ of PRID is used as the query set while $cam2$ is the gallery set and vice-versa. R1 and R10 indicates rank-1 and rank-10 identification accuracy respectively.

[36, 25]. The comparison indicates the effectiveness of the proposed SC-IMGAN in a cross-domain setting.

4.4. Ablation Study

Effectiveness of Improved Quartet Loss

In this section, we investigate the effectiveness of the improved quartet loss. We train the Re-ID model with both the triplet loss and quartet loss to evaluate the effectiveness of improved quartet loss. From Table 3, we see that the rank-1 accuracy of $ImpQuartet$ outperforms $Quartet$ by 6.3%, 3.2%, 1.6%, 3.2%, 1.2% and 6.1% for CUHK01, CUHK03, VIPeR, PRID2011, iLIDS and Market-1501 datasets; indicating that considering the distance between the positive pairs within the loss is important. Further, we show that the proposed multi-component image generation helps to improve performance compared to CycleGAN and StarGAN.

Effectiveness of Identity Mapping Loss

For person Re-ID, the style transferred image of a person should have the same identity before and after image translation. To this end, our proposed identity mapping multi-component network forces the generator to preserve the identity of the real images such that after style adaptation the identity will be the same. To justify the effectiveness of the identity mapping loss, we compare a multi-component model without the identity mapping loss ($MC-GAN$), with a multi-component GAN with the identity mapping loss but without the semantic-consistency loss ($IMGAN$), as shown in Figure 4 and Table 3. From Table 3, rank-1 performance of IMGAN achieves a 0.9%, 1.6%, 0.4%, 0.7%, 0.6% and 1.5% increase over MC-GAN on CUHK01, CUHK03, VIPeR, PRID2011, iLIDS and Market-1501 datasets. We also observe from Figure 4 that without the identity loss, it is harder for the model to generate images with the same identity, reducing performance.

Effectiveness of Semantic Consistency Loss

In person Re-ID, domains are vastly different such as between the PRID2011 and Market-1501 datasets. This leads to a performance drop when transforming the images from one domain to others. Thus, preserving semantic information across domains helps improve translation across domains, and improves performance. To evaluate the effectiveness of the proposed feature-level loss, we evaluate a multi-component model for target domain style adaptation without the semantic-consistency loss ($IMGAN$), and compare this to the proposed approach (see Figure 4 and Table 3). According to Table 3, removing the the semantic-consistency loss drops performance by 0.7%, 1.1%, 0.9%, 0.6%, 0.5% and 1.1% on the CUHK01, CUHK03, VIPeR, PRID2011, iLIDS and Market-1501 datasets, indicating that the semantic consistency loss leads to a substantial improvement in performance.

Method	C1	C3	V	P	L	M
	R1	R1	R1	R1	R1	R1
Triplet	79.5	83.8	61.0	71.0	71.5	79.3
Quartet	83.9	85.5	68.7	75.0	82.4	83.6
Imp Quartet	90.2	88.7	70.3	78.2	83.6	89.7
CycleGAN + Imp Quartet	90.8	89.0	70.9	78.5	84.0	90.2
StarGAN + Imp Quartet	91.4	89.5	71.2	78.8	84.1	91.6
MC-GAN + Imp Quartet	93.7	90.1	71.5	79.4	84.9	92.1
IMGAN + Imp Quartet	94.6	91.7	71.9	80.1	85.5	93.6
SC-IMGAN + Imp Quartet	**95.3**	**92.8**	**72.8**	**80.7**	**86.0**	**94.7**

Table 3: Ablation studies on CUHK01, CUHK03, VIPeR, PRID2011, iLIDS and Market-1501 datasets. "MC-GAN + Imp Quartet": multi-component image generation network without identity-mapping loss and semantic-consistency loss but with improved Quartet loss. "IMGAN + Imp Quartet": with identity mapping loss and improved Quartet loss. "SC-IMGAN + Imp Quartet": with semantic-consistency and identity mapping loss and improved Quartet loss. We also compared to StarGAN [9] and CycleGAN [53]. R1 indicates rank-1 identification accuracy.

From the above, we see that both the semantic consistency loss and identity mapping loss improve performance. It is clear that person Re-ID also benefits from being trained with multiple different styles of a person. Inspecting Figure 4, the proposed SC-IMGAN helps preserve the identity and semantics during image domain translation. The proposed model overcomes the limitation of CycleGAN only being able to transfer between two domains by transferring styles among N domains, greatly reducing the computational requirements over CycleGAN. SC-IMGAN also performs better than StarGAN which uses an external code for image translation, and thus can not perform well when domains are vastly different.

From Table 1 and 3, it can be observed that the improved quartet loss outperforms [18]. While [18] used a quartet of images, it did not take into account the distance between the positive pairs and therefore features from the same identity may be separated by a large intra-class distance which leads to a drop in performance. The proposed improved quartet overcomes this problem by not only increasing the inter-class distance with multiple different probe images, but also simultaneously penalising large intra-class distances.

To obtain further insight, a t-SNE visualisation of learned embeddings for the PRID2011 dataset is shown in Figure 5. Here, we have considered 10 classes for clear visualization. From Figure 5, the proposed improved quar-

Figure 4: Examples of style-transferred images in six domains : (a) the samples of the proposed SC-IMGAN (with identity mapping and semantic-consistency loss), (b) samples without semantic-consistency loss, (c) samples without identity mapping loss, (d) StarGAN and (e) CycleGAN. The six rows represent real images from the CUHK01, CUHK03, VIPeR, PRID2011, iLIDS and Market-1501 domains. Inputs from all these domains are transferred to the distributions of CUHK01, CUHK03, VIPeR, PRID2011, iLIDS and Market-1501 respectively, represented as $D1 - D6$.

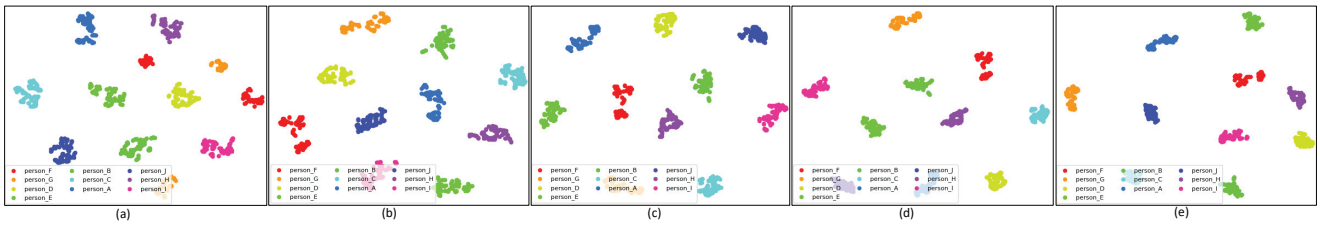

Figure 5: t-SNE visualizations of the CNN activations for PRID2011 (a) quartet, (b) improved quartet, (c) MC-GAN, (d) IMGAN and (e) SC-IMGAN. The 10 different colours correspond 10 different identities. It can be seen that SC-IMGAN has the tightest grouping of points, indicating that it is best suited to separating the classes.

tet loss optimizes the embedding space such that the data points with the same identity are closer to each other than we see for the quartet loss, for example, the red points are wrongly clustered in Figure 5 (a). The effectiveness of the identity mapping loss and semantic loss can also be clearly observed.

5. Conclusion

In this paper, we propose SC-IMGAN, a semantic-consistency and identity mapping multi-component GAN for deep person Re-ID. The SC-IMGAN model learns to transfer styles among six domains to generate new training images. An identity mapping loss ensures that style-transferred images contain the same identity as the original images, and a semantic-consistency loss is also proposed to ensure that encoders extract the same high level features for images belonging to the same identity, regardless of the image domains; improving performance gain during cross-domain translation. The style-transferred images are then used with real images to train the proposed four-stream person Re-ID network. To ensure that we maximise the distance between negative pairs relative to the positive pairs and with respect to multiple different probe images, we propose to use an improved quartet loss with joint verification-identification, which further keeps intra-class features close to each other. Evaluations on 6 popular databases show our technique outperforms current state-of-the-art methods. To emulate a real-world scenario, we performed cross-domain experiments on the proposed methods and the results demonstrate our architecture superior performance in this challenging real world scenario.

Acknowledgements

This research was supported by the Australian Research Council's Linkage Project "Improving Productivity and Efficiency of Australian Airports" (140100282). The authors would also like to thank QUT High Performance Computing (HPC) for providing the computational resources for this research.

References

[1] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, 2017.

[2] S. Bak and P. Carr. One-shot metric learning for person re-identification. In *CVPR*, 2017.

[3] S. Bak, P. Carr, and J.-F. Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *ECCV*, 2018.

[4] X. Chang, T. M. Hospedales, and T. Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, 2018.

[5] S. Z. Chen, C. C. Guo, and J. H. Lai. Deep ranking for person re-identification via joint representation learning. *IEEE TIP*, 2016.

[6] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, 2017.

[7] W. Chen, X. Chen, J. Zhang, and K. Huang. A multi-task deep network for person re-identification. In *AAAI*, 2017.

[8] 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. In *CVPR*, 2016.

[9] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[11] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *PR*, 2015.

[12] H. Fan, L. Zheng, C. Yan, and Y. Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM TOMCCAP*, 2018.

[13] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*. 2014.

[15] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[16] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011.

[17] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[18] A. Khatun, S. Denman, S. Sridharan, and C. Fookes. A deep four-stream siamese convolutional neural network with joint verification and identification loss for person re-detection. In *WACV*, 2018.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*. 2012.

[20] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017.

[21] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013.

[22] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[23] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, 2019.

[24] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *IEEE TIP*, 2017.

[25] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang. Adaptive transfer network for cross-domain person re-identification. In *CVPR*, 2019.

[26] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *NIPS*. 2016.

[27] T. Matsukawa and E. Suzuki. Person re-identification using cnn features learned from combination of attributes. In *ICPR*, 2016.

[28] M. M. Rahman, C. Fookes, M. Baktashmotlagh, and S. Sridharan. Multi-component image translation for deep domain generalization. In *WACV*, 2019.

[29] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*, 2016.

[30] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. *ICLR*, 2017.

[31] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016.

[32] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016.

[33] L. Wang, W. Ouyang, X. Wang, and H. Lu. Visual tracking with fully convolutional networks. In *ICCV*, 2015.

[34] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, 2014.

[35] Y. Wang, L. Wang, Y. You, X. U. Zou, V. Chen, S. P. Li, G. Huang, B. Hariharan, and K. Q. Weinberger. Resource aware person re-identification across multiple resolutions. In *CVPR*, 2018.

[36] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.

[37] L. Wu, C. Shen, and A. van den Hengel. Personnet: Person re-identification with deep convolutional neural networks. *CoRR*, 2016.

[38] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016.

[39] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, 2018.

[40] Q. Yang, H.-X. Yu, A. Wu, and W.-S. Zheng. Patch-based discriminative feature learning for unsupervised person re-identification. In *CVPR*, 2019.

[41] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *ICPR*, 2014.

[42] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017.

[43] D. Yoo, N. Kim, S. Park, A. S. Paek, and I.-S. Kweon. Pixel-level domain transfer. In *ECCV*, 2016.

[44] H.-X. Yu, A. Wu, and W.-S. Zheng. Unsupervised person re-identification by deep asymmetric metric embedding. *TPAMI*, 2019.

[45] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. CVPR, 2017.

[46] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[47] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned CNN embedding for person re-identification. *CoRR*, 2016.

[48] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[49] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. *CVPR*, 2017.

[50] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018.

[51] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camstyle: A novel data augmentation method for person re-identification. *IEEE TIP*, 2019.

[52] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng. Point to set similarity based deep feature learning for person re-identification. In *CVPR*, 2017.

[53] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.