Cross-Modality Person Re-Identification with Generative Adversarial Training

Pingyang Dai[1,2], **Rongrong Ji**[1,2*], **Haibin Wang**[1,2], **Qiong Wu**[2], **Yuyu Huang**[1,2]
[1] Fujian Key Laboratory of Sensing and Computing for Smart City, Xiamen University, China
[2] School of Information Science and Engineering, Xiamen University, China
{pydai, rrji}@xmu.edu.cn, {haibin, qiong, huangyuyu}@stu.xmu.edu.cn

Abstract

Person re-identification (Re-ID) is an important task in video surveillance which automatically searches and identifies people across different cameras. Despite the extensive Re-ID progress in RGB cameras, few works have studied the Re-ID between infrared and RGB images, which is essentially a *cross-modality* problem and widely encountered in real-world scenarios. The key challenge lies in two folds, *i.e.*, the lack of discriminative information to re-identify the same person between RGB and infrared modalities, and the difficulty to learn a robust metric for such a large-scale cross-modality retrieval. In this paper, we tackle the above two challenges by proposing a novel cross-modality generative adversarial network (termed *cm*GAN). To handle the lack of insufficient discriminative information, we design a cutting-edge generative adversarial training based discriminator to learn discriminative feature representation from different modalities. To handle the issue of large-scale cross-modality metric learning, we integrate both identification loss and cross-modality triplet loss, which minimize inter-class ambiguity while maximizing cross-modality similarity among instances. The entire *cm*GAN can be trained in an end-to-end manner by using standard deep neural network framework. We have quantized the performance of our work in the newly-released SYSU RGB-IR Re-ID benchmark, and have reported superior performance, *i.e.*, Cumulative Match Characteristic curve (CMC) and Mean Average Precision (MAP), over the state-of-the-art works [Wu *et al.*, 2017], at least 12.17% and 11.85% respectively.

1 Introduction

Person re-identification (Re-ID) has received ever-increasing research focus recently, which aims to match pedestrian images across different cameras. The key challenge lies in that different cameras usually suffer from significant changes in different views, human body poses, illumination and backgrounds. To this end, the existing works can be roughtly categorized into either metric learning methods or feature learning methods. The former targets at learning a distance metric to handle the above robust matching problem [Zheng *et al.*, 2011; Köstinger *et al.*, 2012; Wang *et al.*, 2013; Li *et al.*, 2013; Xiong *et al.*, 2014; Liao *et al.*, 2015; Paisitkriangkrai *et al.*, 2015; Liao and Li, 2015; Wang *et al.*, 2016; Subramaniam *et al.*, 2016]. The latter is to learn invariant feature directly, upon which efficient L2 or Cosine distance can be applied [Farenzena *et al.*, 2010; Kviatkovsky *et al.*, 2013; Zhao *et al.*, 2013; Yang *et al.*, 2014; Liao *et al.*, 2015; Matsukawa *et al.*, 2016; Ma *et al.*, 2017].

Figure 1: Examples of RGB images and infrared (IR) images in SYSU RGB-IR Re-ID dataset. The 1^{st},2^{nd},4^{th},5^{th}rows are RGB images, while 3^{rd},6^{th} rows are IR images.

However, very few works have paid attention to the Re-ID between RGB cameras and infrared cameras, which, in our consideration, is essentially a *cross-modality* problem and widely encountered in real-world scenarios. In many applications, the surveillance cameras could be heterogeneous, such as near-infrared(IR), thermal and depth cameras. Especially, many new-generation surveillance cameras nowadays support automatic switching between RGB and infrared modalities, which facilitates such cameras to work at night. The key challenges lies in two folds, *i.e.*, the lack of discriminative information to re-identify the same person between RGB and infrared modalities, and the difficulty to learn a robust metric towards for such a large-scale cross-modality retrieval. Very recently, the work in [Wu *et al.*, 2017] introduces a cross-modality RGB-IR dataset named SYSU RGB-IR Re-ID, as

*Corresponding author: Rongrong Ji (rrji@xmu.edu.cn)

shown in Fig.1. The proposed method analysizes three different network structures and uses deep zero padding for evolving domain-specific structure in one-stream network, which is suitable for RGB-IR Re-ID. However, two serious challenging problems in person Re-ID remain open. First, learning a distance metric is difficult, especially in large-scale cross-modality retrieval. The second challenge is the unbalanced data from different modalities which indicates inconsistent distribution and representation of different modalities and heterogeneous gap.

In this paper, we tackle the above two challenges by proposing a novel cross-modality generative adversarial network (termed *cm*GAN). To handle the issue of insufficient discriminative information, we leverage the cutting-edge generative adversarial training theory to design our own discriminator to learn discriminative feature representation from different modalities. To handle the issue of the unbalanced data from different modalities, we proposed a novel hybrid loss that integrates identification loss and cross-modality triplet loss, which minimize inter-class ambiguity while maximizing cross-modality similarity among instances . In particular, the proposed *cm*GAN consists of a deep convolutional neural network as generator to generate modality-invariant representation for RGB and IR images in a common subspace, as well as a modality classifier as discriminator that discriminates between different modalities. The generator is supervised and optimized by identification loss and cross-modality triplet loss, by which identification loss can separate inter-class embedding while cross-modality triplet loss minimizes the gap among RGB and infrared representations. The discriminator is a modality classifier that tries to discriminate RGB-IR image representations between different modalities. In the proposed framework, the generator and discriminator beat each other as a minimax game to learn discriminative common representation. Through the joint exploitation of the above, heterogeneous data can be directly compared by such a robust distance metric for cross-modality person Re-ID. The main contributions of this paper are three-fold:

- We explore the problem of cross-modality RGB-IR person Re-ID by exploiting the cross-modality representation learning from the perspective of generative adversarial training. To the best of our knowledge, it is the first effort towards GAN based cross-modality person Re-ID.
- We design a loss function for cross-modality generative adversarial network (cmGAN) to learn discriminative common representations. With this loss function, the generator and discriminator in *cm*GAN beat each other with minimax game to learn cross-modality representations.
- Extensive experiments on challenging RGB-IR re-identification dataset demonstrate the advantages of our proposal over state-of-the-art methods of CMC by 12.17% in terms of mAP by 11.85%.

The rest of paper is organized as follows. Sec.2 outlines related works on person Re-ID, cross modality retrieval and generative adversarial networks. Sec.3 describes the proposed framework for learning discriminative representations for RGB-IR person Re-ID. Experimental results of our approach are demonstrated in Sec.4. Finally we conclude this paper in Sec.5.

2 Related Work

2.1 Person Re-ID

There are two fundamental components in Person Re-ID: distance metric learning and feature representation. For the former group of works, metric learning is to formalize the problem as a supervised metric learning where a projection matrix is sought out. Subsequently, metrics like Mahalanobis metric are well exploited. [Köstinger *et al.*, 2012] proposed a large-scale metric learning from equivalence constraint (KISSME), which essentially models a log likelihood ratio test between two Gaussian distributions. [Ding *et al.*, 2015] proposed a deep image representation based on relative distance comparison for person Re-ID. Another group is to learn discriminative features that can be efficiently combined by existing L2 or Cosine distances. In particular, several deep learning schemes are exploited. [Yi *et al.*, 2014] constructed a Siamese neural network to learn pairwise similarity and used pedestrian body parts to train their CNN model. [Ahmed *et al.*, 2015] proposed a new patch matching layer that compares the activation of two images in neighboring pixels. However, the aforementioned works only address the person Re-ID problem from the same modality. They cannot handle multimodal data with consistent distribution, which cannot bridge the heterogeneous gap.

2.2 Cross Modality Retrieval

Cross-modality retrieval [Peng *et al.*, 2017a]targets at search instances across different modality data, such as searching text in image database related to it semantically. Search between image and document is a representative cross modality retrieval, which has attracted extensive research focus in the past decade. Representative methods include, but not limited to, traditional statistical correlation analysis, DNN-based methods [Zhang *et al.*, 2014], cross-media graph regularization methods, metric learning methods, and dictionary learning methods. The widely-used cross-modality retrieval datasets include Wikipedia, XMedia, NUS-WIDE and Pascal VOC 2007. Among all the methods, the implementation of deep neural network has a significant influence on retrieval performance because of its effectiveness and efficiency. And the recent endeavors also facilitate the learning of cross-modality binary codes, such as [Srivastava and Salakhutdinov, 2014]. A deep semantic hashing with generative adversarial networks is proposed by [Qiu *et al.*, 2017], which explores semi-supervised GAN to generate synthetic training data for hashing. Note that, our scenario is related to the above works, which however differs in person Re-ID. Therefore such works cannot be directly applied.

2.3 Generative Adversarial Network

After first proposed by [Goodfellow *et al.*, 2014], GAN has received ever-increasing research focus in computer vision and artificial intelligence research [Arjovsky *et al.*, 2017; Isola *et al.*, 2017; Ledig *et al.*, 2017]. The GAN consists

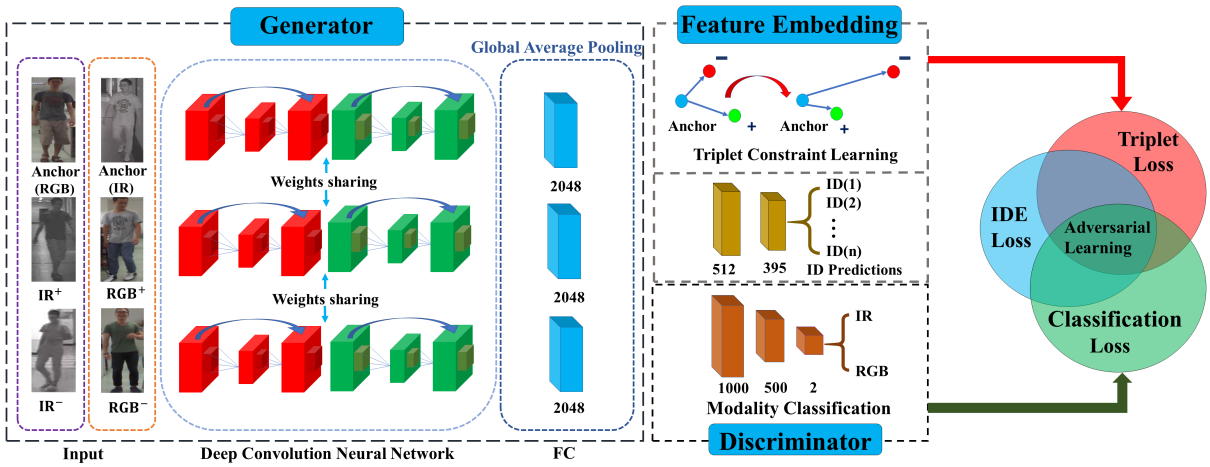

Figure 2: The proposed *cm*GAN framework. It consists of the two components. A deep convolutional neural network as generator with identification loss and cross-modality triplet loss to generate modality-invariant representation for RGB and IR images in common subspace, as well as a modality classifier as discriminator that discriminates between different modalities. The generator and discriminator beat each other as a minimax game to learn discriminative common representation for person Re-ID.

of a generator and a discriminator, the former is to capture the distribution of real data, while the latter reveals whether a sample is fake or real.

To the best of our knowledge, the existing GAN methods typically handle intra-modality sample generation. As among few pioneering works, [Zhao *et al.*, 2017] propose a deep modal that turns cross-view image hashing into single-view GAN for cross modality image retrieval. [Peng *et al.*, 2017b] proposed a method to effectively correlate large-scale heterogeneous data by GAN. [Wang *et al.*, 2017] presented an adversarial cross-modal retrieval method to seek an effective common subspace based on adversarial learning.

3 The Proposed *cm*GAN

3.1 Problem Formulation

Formally speaking, let **V** be the RGB images and **I** be the IR images or thermal images. The multi-modal dataset is represented as $D = \{D_{tr}, D_{te}\}$, where $D_{tr} = \{V_{tr}, I_{tr}\}$ denotes the training data and D_{te} denotes the testing data , in which $D_{te} = \{V_{te}, I_{te}\}$. Here, $v_j \in \mathbb{R}^d$ is the feature vector extracted from the RGB image and $i_j \in \mathbb{R}^d$ is the feature extracted from the IR image, where d is the feature dimension. Assume D_{tr} contains a set of n training images with the corresponding identity labels as $Y = \{y_i\}_1^n$.

3.2 The Proposed Model

Fig.2 shows the framework of the proposed cmGAN model, which contains generator and discriminator components. Features **V** and **I** extracted from RGB and IR modalities respectively are fed into the fully-connected layers to generate a 2,048 dimension feature. Then, the generator is trained with both the identification loss and cross-modality triplet loss (detailed in **Generator**). Finally, the modality classifier plays a role as a discriminator to discriminate between different modalities (detailed in **Discriminator**). The model components are introduced in details as below:

Generator We construct the generative model with two kinds of loss functions *i.e.*, the intra-class feature embedding loss and the inter-modality loss. The first loss is to ensure the intra-modal discrimination. In order to project feature representation for each modality in the common subspace, several fully-connected layers are built with the identification loss. The feed-forward network is activated by softmax and added after the global average pooling, which outputs a probability distribution of person identifications. This intra-class feature embedding loss by uses the following probability distributions as:

$$\ell_{ide}(\theta_{ide}) = -\frac{1}{M}\sum_{j=1}^{M}(y_i \cdot (logp_j(v_j) + logp_i(i_j)). \quad (1)$$

Here, θ_{ide} denotes the parameters of the feed-forward network and p_j is the probability distribution of each ID. The second loss, *i.e.* the cross-modality loss is to impose triplet constraints to minimize the gap among features of identical persons from different modalities. Formally speaking, let $D_i = \{v_a, i_p, i_n\}$ or $D_i = \{i_a, v_p, v_n\}$ be the input feature of different modalities. Such D_i forms the i-th triple, where (v_a, i_p) and (i_a, v_p) are from the same person in different modalities, while i_n and v_n have different identification labels respectively. We then compute the inter-modality loss respectively:

$$\ell_{triplet}\mathcal{V}(\theta_V) = \sum_{(a,p,n)}[D_{a,p}^2 - D_{a,n}^2 + \xi]_+, \quad (2)$$

$$\ell_{triplet}\mathcal{I}(\theta_I) = \sum_{(a,p,n)}[D_{a,p}^2 - D_{a,n}^2 + \xi]_+. \quad (3)$$

Here, $[z]_+ = max(z, 0)$, the threshold ξ is a margin that is enforced between positive and negative pairs. The overall cross-modality triplet loss can be formulated as a combination of $\ell_{triplet}\mathcal{V}(\theta_V)$ and $\ell_{triplet}\mathcal{I}(\theta_I)$.

$$\ell_{triplet}(\theta_V, \theta_I) = \ell_{triplet}\mathcal{V}(\theta_V) + \ell_{triplet}\mathcal{I}(\theta_I). \quad (4)$$

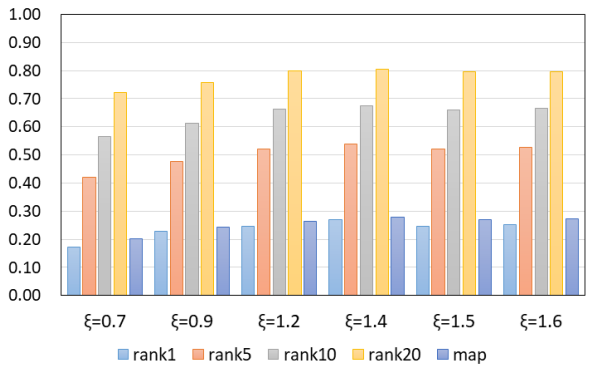

Figure 3: Performance of cmGAN with different values of margin ξ in triplet loss funciton.

By combining the intra class embedding loss with the inter-modality loss, we come up with the final objective function as:

$$\ell_{gen}(\theta_V, \theta_I, \theta_{ide}) = \alpha \cdot \ell_{triplet} + \beta \cdot \ell_{ide}, \tag{5}$$

where α and β are hyper-parameters.

Discriminator We construct a modality classifier D with parameters θ_D to act as discriminator. This modality classifier is to learn to judge whether a learned vector of representation is within the same modality or between different modalities, which acts as an adversary. The modality classifier is consisted of a 3-layer feed-forward neural network, upon which the discriminator (adversarial) loss is defined by the cross-entropy loss of the modality classifier as follows:

$$\ell_{dis}(\theta_D) = -\frac{1}{M}\sum_{j=1}^{M}(m_i \cdot (logD(v_j;\theta_D) + logD(i_j;\theta_D))), \tag{6}$$

where m_i is the modality label, and $D(\cdot;\theta_D)$ is the modality probability of each input image.

3.3 The Training Algorithm

The model learning is conducted end-to-end by jointly minimaxing generator and the discriminator losses , which is similar to [Peng *et al.*, 2017b] and [Wang *et al.*, 2017]. Since the generator and the discriminator run as a minimax game that targets at beating each other, the optimization carries out two sub-processes:

$$\widehat{\theta}_V, \widehat{\theta}_I, \widehat{\theta}_{ide} = \arg\min_{\theta_V, \theta_I, \theta_{ide}} (\ell_{gen}(\theta_V, \theta_I, \theta_{ide}) - \gamma * \ell_{dis}(\widehat{\theta}_D)), \tag{7}$$

$$\widehat{\theta}_D = \arg\max_{\theta_D}(\ell_{gen}(\widehat{\theta}_V, \widehat{\theta}_I, \widehat{\theta}_{ide}) - \gamma * \ell_{dis}(\theta_D)). \tag{8}$$

Here, γ is a hyper-parameter. The generative model is trained for K steps in each iteration in the training stage. The overall training procedure is shown in Algorithm.1.

Algorithm 1 The Learning of the Proposed cmGAN model

Require:
Training dataset $\{x_1,y_1,s_1\},\{x_2,y_2,s_2\},...$
(x is raw image, y is Id label, s is domain label),
A per-trained CNN model $v = f(x;\theta_1)$ as generator,
A CNN model $p = d(v;\theta_2)$ as discriminator,
parameter: α, β, γ, η, K, ξ.

Repeat until convergence:
for each $i \in labels$ **do**
  randomly create the pairs:$(\{x_a,y_i,s\},\{x_b,y_i,\bar{s}\})$
end for
Calculate the generator loss
$\mathcal{V} = f(\mathcal{X};\theta_1)$
$\ell_{triplet} = \sum_{a,p,n}[D^2_{a,p} - D^2_{a,n} + \xi]_+$
$(y_a = y_p, y_a \neq y_n, s_a \neq s_p, s_a \neq s_n)$
$\ell_{ide} = \text{CrossEntropy\_Loss}(\mathcal{V}, \mathcal{Y})$
$\ell_{gen} = \alpha * \ell_{triplet} + \beta * \ell_{ide}$
Calculate the discriminator loss
$\mathcal{P} = d(\mathcal{V};\theta_2)$
$\ell_{dis} = \text{CrossEntropy\_Loss}(\mathcal{P}, \mathcal{S})$
Train the parameters θ_1 and θ_2
if $e \,\%\, K \neq 0$ **then**
  $\ell_{Total} = \ell_{gen} - \gamma * \ell_{dis}$;
else
  $\ell_{Total} = \eta * (\gamma * \ell_{dis} - \ell_{gen})$;
end if
Update ℓ_{Total}
return learned representation in common space $f(x;\theta_1)$

4 Experiments

In this section, we conduct extensive experiments to evaluate the efficacy of the proposed method. In the experiments reported blow, we compare our proposed *cm*GAN method with the state-of-the-art methods on SYSU RGB-IR Re-ID dataset [Wu *et al.*, 2017] to verify its effectiveness. Then we conduct further analysis to investigate the performance of *cm*GAN in more details.

4.1 Datasets and Settings

The SYSU RGB-IR Re-ID dataset[1] is the first benchmark for cross-modality (RGB-IR) Re-ID, which is captured by 6 cameras, including two IR cameras and four RGB ones. This dataset contains 491 persons with total 287,628 RGB images and 15,792 IR images from four RGB cameras and two IR cameras. The dataset is separated into the training set and the test set, where images of the same person can only appear in either set. And the training set consists of total 32,451 images including 19,659 RGB images and 12,792 IR images. It is a very challenging dataset due to the great differences between two modalities.

4.2 Evaluation Protocol

Our experiments follow the evaluation protocol in [Wu *et al.*, 2017]. When we evaluate our model, we set RGB images for gallery images, and those IR images are for probe set. There are two modes, *all-search* mode and *indoor-search* mode. For *all-search* mode, RGB cameras 1, 2, 4 and 5 are for gallery set and IR cameras 3 and 6 are for probe set. For *indoor-search* mode, RGB cameras 1 and 2 (excluding outdoor cameras 4 and 5) are for gallery set and IR cameras 3 and 6 are for probe

[1] http://isee.sysu.edu.cn/project/RGBIRReID.htm

Method	All-search								Indoor-search							
	Single-shot				Multi-shot				Single-shot				Multi-shot			
	r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP
HOG+Euclidean	2.76	18.25	31.91	4.24	3.82	22.77	37.63	2.16	3.22	24.68	44.52	7.25	4.75	29.06	49.38	3.51
HOG+CRAFT	2.59	17.93	31.50	4.24	3.58	22.90	38.59	2.06	3.03	24.07	42.89	7.07	4.16	27.75	47.16	3.17
HOG+CCA	2.74	18.91	32.51	4.28	3.25	21.82	36.51	2.04	4.38	29.96	50.43	8.70	4.62	34.22	56.28	3.87
HOG+LFDA	2.33	18.58	33.38	4.35	3.82	20.48	35.84	2.20	2.44	24.13	45.50	6.87	3.42	25.27	45.11	3.19
LOMO+CCA	2.42	18.22	32.45	4.19	2.63	19.68	34.82	2.15	4.11	30.60	52.54	8.83	4.86	34.40	57.30	4.47
LOMO+CRAFT	2.34	18.70	32.93	4.22	3.03	21.70	37.05	2.13	3.89	27.55	48.16	8.37	2.45	20.20	38.15	2.69
LOMO+CDFE	3.64	23.18	37.28	4.53	4.70	28.23	43.05	2.28	5.75	34.35	54.90	10.19	7.36	40.38	60.33	5.64
LOMO+LFDA	2.98	21.11	35.36	4.81	3.86	24.01	40.54	2.61	4.81	32.16	52.50	9.56	6.27	36.29	58.11	5.15
Asymmetric FC layer network [Wu *et al.*, 2017]	9.30	43.26	60.38	10.82	13.06	52.11	69.52	6.68	14.59	57.94	78.68	20.33	20.09	69.37	85.08	13.04
Two-stream network [Wu *et al.*, 2017]	11.65	47.99	65.50	12.85	16.33	58.35	74.46	8.03	15.60	61.18	81.02	21.49	22.49	72.22	88.61	13.92
One-stream network [Wu *et al.*, 2017]	12.04	49.68	66.74	13.67	16.26	58.14	75.05	8.59	16.94	63.55	82.10	22.95	22.62	71.74	87.82	15.04
One-stream network (zero-padding) [Wu et al., 2017]	14.80	54.12	71.33	15.95	19.13	61.40	78.41	10.89	20.58	68.38	85.79	26.92	24.43	75.86	91.32	18.64
cmGAN (proposed) (only with triplet loss)	18.37	64.12	**80.96**	22.04	19.52	**73.33**	**89.93**	**31.38**	22.29	69.63	84.80	15.56	25.29	77.57	91.72	22.49
cmGAN (proposed) (only with ID predictions)	11.52	41.72	56.54	13.03	17.23	52.77	68.14	25.46	13.23	44.87	59.65	8.65	21.40	58.59	71.42	17.00
cmGAN (proposed)	**26.97**	**67.51**	80.56	**27.80**	**31.49**	72.74	85.01	22.27	**31.63**	**77.23**	**89.18**	**42.19**	**37.00**	**80.94**	**92.11**	**32.76**

Table 1: Performance under all-search and indoor-search mode. r1, r10, r 20 denote rank-1, 10, 20 accuracies (%)

set which is easier than former. For both modes, *single-shot* and *multi-shot* settings are used. we randomly choose one image and ten images of the identity to form the gallery set for *single-shot* and *multi-shot* setting in RGB images respectively. Given a probe image from camera 3 or 6, we compute similarities between the probe image and gallery images. It is important to note the probe image and gallery image are from different locations, that is, camera 2 and camera 3 are in the same location, so probe images of camera 3 skip the gallery images of camera 2. After computing similarities, we can get a ranking list according to descending order of similarities.

4.3 Implementation Details

We use NVIDIA GeForce 1080Ti graphics cards for our entire experiments. And we extract 2,048d features for RGB and IR images from the global average pooling layer in ResNet-50. Three fully-connected layers are used in the modality classifier. Furthermore, We stick to the two fully-connected layers for ID predictions.

The batch size is set to 20. Empirically after testing several groups of parameter combinations, the generative model training step K is set to be 5. The adaptive parameter γ controls the weight of the discriminator loss and is fixed to be 0.05. We set margin ξ in cross-modality triplet loss in the range [0.7, 0.9, 1.2, 1.4, 1.5, 1.6] and conduct experiments by fixing other hyper-parameters to evaluate our model. The impact of the ξ is shown in Fig.3. We observe that our method achieves better performance when the margin ξ raising. The best performance reported result of our method are obtained when ξ is set to be 1.4.

Considering the different convergence rate of generator and discriminator, the learning rate (lr) of them are set being different respectively. We set the learning rate of generator range in [0.0001, 0.001, 0.01] and discriminator learning rate also range in [0.0001,0.001,0.01]. After testing several groups of parameter combinations empirically, We get the best result when generator lr is set to be 0.0001 with discriminator lr is set to be 0.001 respectively. Therefore, we set the training epoch to 2000 and more than it.

Empirically, we set α and β by 1:1 proportion because we consider cross-modality triplet loss has an equal effect on the retrieval results as well as identification loss.

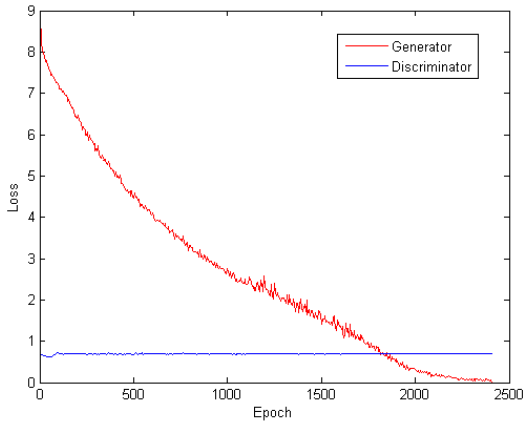

Figure 4: Our observation of generative loss and discriminative loss during the training process.

4.4 Quantitative Evaluations

We shows the comparative results of *cm*GAN against 12 existing methods [Dalal and Triggs, 2005; Liao *et al.*, 2015; Wu *et al.*, 2017] on SYSU RGB-IR Re-ID in Tab.1, including the rank-1, 10, 20 accuracies of Cumulative Match Characteristic curve (CMC) and mean average precision(mAP).

From Tab.1, it is clear that our *cm*GAN method significantly outperforms all existing methods in all categories, which is in terms of *all search* and *indoor search* mode or *one-shot* and *multi-shot* settings. Specifically, the proposed *cm*GAN method outperforms the 2nd best method of [Wu *et al.*, 2017] by using one-stream network (deep zero-padding) on *all search* and *indoor search* tasks in terms of the rank-1 metric by 12.17% (26.97-14.80) and 11.05% (31.63-20.58) under *one-shot* respectively.

In Tab.1, the results of three rows on the bottom show the performance of *cm*GAN and its two variations. It is shown that both the cross-modality triplet loss constraint and identification loss terms contribute to the final retrieval results, which demonstrates that to optimize the proposed model with two loss function together is better than only one of them. We

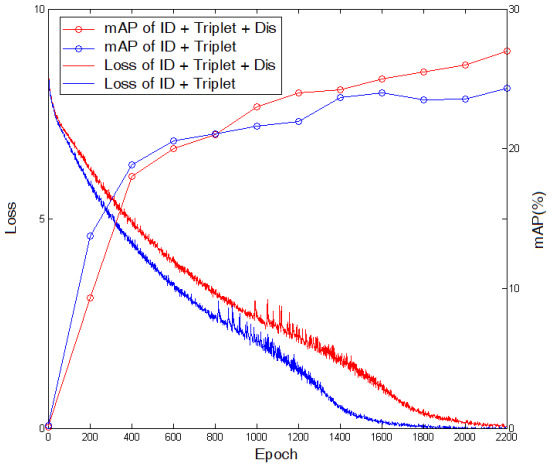

Figure 5: The trend of total loss and total loss without discriminator loss during the training process.

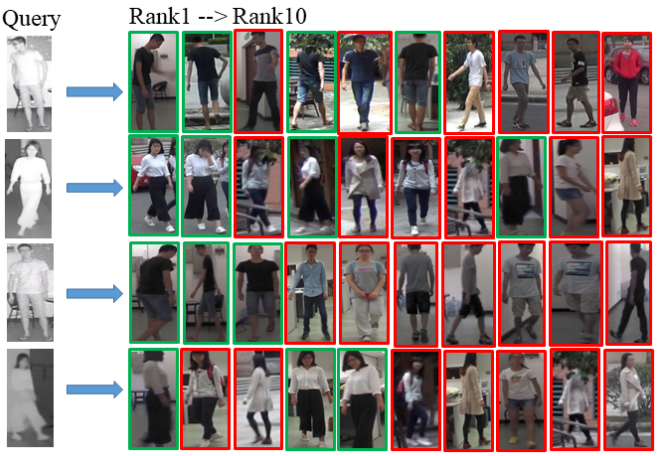

Figure 6: Some examples of retrieval results on SYSU RGB-IR Re-ID using the proposed method.

can also find that the triplet loss constraint contributes more to the performance than the identification term. That means the identification loss do not include cross-modality information, so the cross-modality triplet loss can do a better job in projecting two different modalities features into the same feature subspace.

4.5 Further Analysis on *cm*GAN

In the proposed *cm*GAN, we target at optimizing the proposed method with the generative loss and discriminative loss in adversarial training process. We investigate the values of the generative loss and the discriminative loss from epoch 1 to 2,500, as shown in Fig.4. The figure demonstrates that the generative loss decreases almost monotonously and converges smoothly, when the discriminative loss changes suddenly at first and then stabilizes at about 0.7.

The results in Fig.5 demonstrate that the mAP score keeps increasing and holds until the generative loss converges. We can see the loss of *cm*GAN with the discriminator decreases more slowly than without the discriminator. It means that the generator and the discriminator form an adversary training process to project the feature representations from different modalities into common subspace. We set a small learning rate in order to achieve better performance with the triplet loss convergence. When the triplet loss is lower than 0.1, we consider that triplet loss has converged. Correspondingly, we increase the learning rate to speed up the convergence of the identification loss. Depending on the discriminator, the *cm*GAN has a more smooth convergence process. The results in Fig.5 also validate that the discriminative loss affects the *cm*GAN as a directional guide during the feature embedding into the cross-modality common subspace.

As the results shown in Fig.6, we randomly select some IR images as queries to search from RGB images. The images in the first column are the query images. The retrieved images are sorted from left to right according to descending order of the similarity score. The first two rows are results under *all-search* mode and the last row are under *indoor-search* mode. The correct matches are in the green rectangles, and the false matching images are in the red rectangles. We can see our method gets great shots under most situations.

5 Conclusions

In this paper, we propose a novel cross-modality generative adversarial network (termed *cm*GAN) to learn discriminative common representations for RGB-IR person re-identification problem, which is formulated as cross-modality problem. In our approach, the cross-modality generative adversarial networks consist of a deep convolutional neural network as generator for learning image representations and a modality classifier as discriminator which tries to discriminate between RGB and infrared image modalities. We use identification loss and cross-modality triplet loss together for generator in our *cm*GAN to handle the large-scale cross-modality metric learning problem. Specifically, The identification loss can separate inter-class embedding meanwhile the cross-modality triplet constraints minimize the gap among the representation from different modality. Both generator and discriminator beat each other with a minimax game and the discriminative common representation is generated by the adversarial learning process for person re-identification. Comprehensive experimental results on the challenging cross-modality person re-identification dataset, SYSU RGB-IR Re-ID, have demonstrated our approach outperforms state-of-the-art methods.

Acknowledgments

This work is supported by the National Key R&D Program (No.2017YFC0113000, and No.2016YFB1001503), Nature Science Foundation of China (No.U1705262, No.61772443, and No.61572410), Post Doctoral Innovative Talent Support Program under Grant BX201600094, China Post-Doctoral Science Foundation under Grant 2017M612134, Scientific Research Project of National Language Committee of China (Grant No. YB135-49), and Nature Science Foundation of Fujian Province, China (No. 2017J01125 and No. 2018J01106).

References

[Ahmed *et al.*, 2015] Ejaz Ahmed, Michael J. Jones, and Tim K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, pages 3908–3916, 2015.

[Arjovsky *et al.*, 2017] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017.

[Dalal and Triggs, 2005] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[Ding *et al.*, 2015] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015.

[Farenzena *et al.*, 2010] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, 2010.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976, 2017.

[Köstinger *et al.*, 2012] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295, 2012.

[Kviatkovsky *et al.*, 2013] Igor Kviatkovsky, Amit Adam, and Ehud Rivlin. Color invariants for person reidentification. *PAMI*, 35(7):1622–1634, 2013.

[Ledig *et al.*, 2017] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 105–114, 2017.

[Li *et al.*, 2013] Zhen Li, Shiyu Chang, Feng Liang, Thomas S. Huang, Liangliang Cao, and John R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, pages 3610–3617, 2013.

[Liao and Li, 2015] Shengcai Liao and Stan Z. Li. Efficient PSD constrained asymmetric metric learning for person re-identification. In *ICCV*, pages 3685–3693, 2015.

[Liao *et al.*, 2015] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015.

[Ma *et al.*, 2017] Xiaolong Ma, Xiatian Zhu, Shaogang Gong, Xudong Xie, Jianming Hu, Kin-Man Lam, and Yisheng Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017.

[Matsukawa *et al.*, 2016] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, pages 1363–1372, 2016.

[Paisitkriangkrai *et al.*, 2015] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, pages 1846–1855, 2015.

[Peng *et al.*, 2017a] Yuxin Peng, Xin Huang, and Yunzhen Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges. *CoRR*, abs/1704.02223, 2017.

[Peng *et al.*, 2017b] Yuxin Peng, Jinwei Qi, and Yuxin Yuan. Cm-gans: Cross-modal generative adversarial networks for common representation learning. *CoRR*, abs/1710.05106, 2017.

[Qiu *et al.*, 2017] Zhaofan Qiu, Yingwei Pan, Ting Yao, and Tao Mei. Deep semantic hashing with generative adversarial networks. In *ACM SIR*, pages 225–234, 2017.

[Srivastava and Salakhutdinov, 2014] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. *Journal of Machine Learning Research*, 15(1):2949–2980, 2014.

[Subramaniam *et al.*, 2016] Arulkumar Subramaniam, Moitreya Chatterjee, and Anurag Mittal. Deep neural networks with inexact matching for person re-identification. In *NIPS*, pages 2667–2675, 2016.

[Wang *et al.*, 2013] Hua Wang, Feiping Nie, and Heng Huang. Multi-view clustering and feature learning via structured sparsity. In *ICML*, pages 352–360, 2013.

[Wang *et al.*, 2016] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by discriminative selection in video ranking. *IEEE TPAMI*, 38(12):2501–2514, 2016.

[Wang *et al.*, 2017] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *ACM MM*, pages 154–162, 2017.

[Wu *et al.*, 2017] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, pages 5390–5399, 2017.

[Xiong *et al.*, 2014] Fei Xiong, Mengran Gou, Octavia I. Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, pages 1–16, 2014.

[Yang *et al.*, 2014] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z. Li. Salient color names for person re-identification. In *ECCV*, pages 536–551, 2014.

[Yi *et al.*, 2014] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Deep metric learning for person re-identification. In *ICPR*, pages 34–39, 2014.

[Zhang *et al.*, 2014] Hanwang Zhang, Yang Yang, Huan-Bo Luan, Shuicheng Yan, and Tat-Seng Chua. Start from scratch: Towards automatically identifying, modeling, and naming visual attributes. In *ACM MM*, pages 187–196, 2014.

[Zhao *et al.*, 2013] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, pages 3586–3593, 2013.

[Zhao *et al.*, 2017] Xin Zhao, Guiguang Ding, Yuchen Guo, Jungong Han, and Yue Gao. TUCH: turning cross-view hashing into single-view hashing via generative adversarial nets. In *IJCAI*, pages 3511–3517, 2017.

[Zheng *et al.*, 2011] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, pages 649–656, 2011.