RGB Infrared Cross Modal Person Re-identification

Chaitra Jambigi
Indian Institute of Science
Department of Computational and Data Sciences
chaitraj@iisc.ac.in

Abstract

The aim of this work is to effectively perform the Re identification task between RGB and IR images. We try to extract the features common in both RGB IR images and in turn align both these images in a common subspace where both can be matched. We use the disentangled representation learning to disentangle the factors of variations present in image data. We disentangle the identity specific components and identity independent components from the RGB and IR data and use the identity specific components at test time to match RGB and IR images. We call these two latent spaces as Structure space and Appearance space and using Generative model, enforce the constraints so as to bring the identity specific information in Appearance space and identity independent information in Structure space. Experiments are performed on both low dimensional image features and high dimensional image data. The results are comparable with the SOTA techniques and motivate us to explore this area of disentangled representation learning.

1 Introduction

Person re-identification is a well known Computer vision retrieval problem. It aims to match a specific person across multiple non overlapping camera views. When the system is presented with a person of interest (query image), person Reid tells whether this person has been observed in another place or time by other camera (Gallery images) [14]. This task has huge importance in surveillance domain owing to the increasing demand of public safety and the widespread use of large camera networks in public places. This is a challenging problem as it has to match two images of same person under intensive appearance changes due to variations in lighting, camera viewpoints, pose changes.

Existing Reid methods deal with RGB images captured by single modality cameras, posing it as a single modality Reid problem or RGB-RGB Person re identification. However in practical surveillance system, it is needed to capture information of a scene at both day and night time, and visible light cameras cannot capture appearance information of a person under poor illumination conditions, i.e. during night time. This limits its applicability in practical surveillance applications. In such cases Infrared (IR) cameras are commonly used in video surveillance systems and most surveillance cameras are able to automatically switch between RGB to IR mode in dark. This leads us to ask a question whether our existing RGB-RGB Reid solutions are capable enough to deal with Infrared Images? This opens up a new path for exploring an interesting and more challenging problem of RGB-IR Person re-identification or Cross Modal Person re identification [12]

RGB and IR images are heterogeneous with very different visual characteristics. Firstly, by imaging principle aspect, the wavelengths of RGB and IR images are different. Secondly, Figure 1 explains these differences where the first row has RGB images having three channels containing information of visible light. The second row has IR images having 1 channel containing information of invisible light. The color information which is the main cue for identity of a person is not present in RGB-IR Person Reid. Thus Cross Modal Reid has 2 main kinds of challenges. One is the Intra-modality variations caused by viewpoint changes, pose variations and illumination changes which is also present in RGB-RGB Person Reid. Second is the Cross modality

variations caused by the modality differences in 2 images being compared. Thus here the system has to learn some features from both RGB and IR images which are independent of color.

The key solution for this problem is to extract identity specific cues from both RGB and IR images and use them for person matching. Some of the identity related information of a person can be the body shape, the attributes like hair, bag, etc., the texture of clothes and bio metric attributes like eyes. Also the images contain other information which are independent of identity like the color information and pose variations. In this problem, color is not an identity cue as it is not present in both modalities. Also pose is not an identity cue as we have data of same identity in different poses.

Figure 1: Top row: Rgb images from SYSU-MM01 dataset, Bottom row: Ir images corresponding to Rgb images

To address the aforementioned issues, we design an approach in which we try to disentangle or to separate out the identity specific cues and the identity independent cues from the RGB and IR images. Our approach uses generative model to decompose each pedestrian image into two latent spaces: **structure space** which encodes the geometry and position/pose related structural information and an **appearance space** which encodes the appearance, color and other finer identity related semantics. The properties captured by the two latent spaces are summarised in Table 1. Thus we use two encoders, a Structure encoder to capture the Structure space attributes and an Appearance encoder to capture the Appearance space attributes. We use a generative model to enforce the constraints needed for effective disentanglement of the image into the latent spaces. We make use of a GAN architecture with a Generator to combine these two latent spaces and reproduce the original image and a Discriminator to classify the generated image as Real/Fake.

Structure pace	Appearance space
Common in RGB, IR-Silhouette, pose background, details like hair, bag, etc.	Color of image,shoes, bag, type of clothes eg. Jeans, shorts,etc. face

Table 1: Description of the information encoded in Structure space and Appearance space

2 Related work

Rgb Ir Person Reid deals with matching the Query image (Infrared image) with the Gallery images (Rgb images). This problem was first published in [12] and they proposed deep zero-padding for evolving domain-specific structure automatically in one-stream network optimised for RGB-IR Re-ID tasks. [13] proposed an end 2 end, dual path and metric learning framework, with a bi-directional, dual constrained top ranking loss. [6] tried to disentangle identity specific and identity independent components, however their method relies on the features learnt from [13] network which may not be totally separable. Few image generation techniques [3] [11] are also explored to create corresponding RGB-IR pair, given an image in one modality and feature learning on paired images. Recently, the area of disentangled representation learning is being explored widely[1] It relies on the fact that data is composed of multiple factors of variations such as viewing conditions, the illumination of the object, angle/pose of the object. Generally the dataset is labelled specifically for the task for which it is collected and this task is supervised by the labels. The model becomes invariant to the hidden factors of variations. Eg. MNIST dataset is labelled for digits however, it also has information about the shape, size, slant, width, height of digit, which often gets discarded during the classification. These uninformative factors of variation needs to be separated from the informative ones, rather than directly discarding as they can be useful in other unsupervised learning approaches. In this work, way we try to make use of this idea of disentangled representation learning to extract the features of our interest.[9]

3 Method

3.1 Problem definition

We represent the Rgb image and Infrared image as $x_{rgb} \in \mathbb{R}^{HxWx3}$ and $x_{ir} \in \mathbb{R}^{HxWx3}$ respectively. Note that the IR image is a 3 channel image which we get in input however there is no different information in rest of the 2 channels. H and W are the height and width of the images. Each image has a corresponding label $y \in \{1, 2, 3....., N\}$ During training, a feature extractor network $\phi(.)$ is trained on the combined Rgb set X_{rgb} and Ir set X_{ir}. During test time, query image from one modality (generally ir) is given as input, and gallery images of other modality (generally rgb) captured from different location cameras are used to match with the query. The Euclidean distance between the query features $\phi(x_{ir})$ and the gallery features $\phi(x_{rgb})$ is calculated and the gallery images are ranked as per their similarity with the query images.

3.2 Framework overview

The Cross modal Reid problem suffers from two main challenges: Intra modality variations caused by camera viewpoint, illumination and pose variations, and the Cross modality variations caused by difference in modality of images. To reduce these two discrepancies a disentanglement strategy is used. As mentioned in the introduction we need to effectively decompose the image into two latent spaces, the Structure space and the Appearance space, i.e. we need to enforce some constraints in the form of losses to extract these spaces. To achieve this as shown in Figure 2 we use two Encoders, the Structure encoder $E_s(x_i)$ which is a shallow network to capture the geometrical/positional structure code s_i and the Appearance encoder which is a deeper Res Net architecture to capture the appearance code a_i. The Generator/Decoder $G(s_i, a_i)$ takes as input the Structure code and Appearance code and generate the image $\widehat{x}_i$. A Discriminator $D(\widehat{x}_i)$ then tries to predict whether $\widehat{x}_i$ is Real or Fake.

3.3 Generative Module

Image generation using GANs has seen a lot of fame in the past few years. GANs can be trained to generate realistic and plausible images even from noise input. We make use of this architecture to generate new image mappings from our rgb and ir images. To be precise, we generate two kinds of image mappings, one is Self identity image generation and the other one is cross identity image generation. By doing these kinds of generations, we try to impose the constraints on the latent spaces, the Structure space and the Appearance space. Also since the structure space should not leak any color information, we pass Grey scale single channel images as input to the Structure encoder.

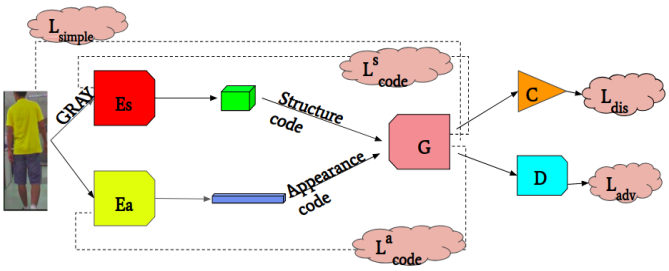
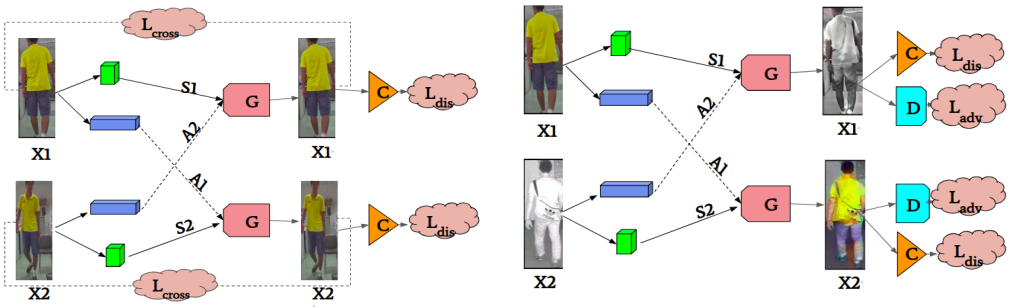


Figure 2: Network architecture and losses involved. The clouds denote the losses involved and the solid shapes denote the sub modules.

Same identity image generation Given an image x_1 of some modality (either rgb or ir) having identity label y_1, we extract s_1 and a_1 i.e. the structure and appearance codes. We first train the generative model to reconstruct the input image itself so as to ensure that the two latent spaces are not throwing away any useful information from the input. This loss acts as a regularisation in total loss. We use pixel wise l_1 loss as,

$$L_{simple} = \mathbb{E} \left\| x_1 - G(a_1, s_1) \right\|_1$$

We assume that the Structure code has geometric information, low level information needed to reconstruct the image back, and Appearance code has more finer identity related high level features.



(a) Same identity, Same modality image generation

(b) Different identity Different modality image generation

Figure 3: Pictorial view denoting the Image generation. Figure a denotes same identity, same modality image generation and Figure b denotes different identity different modality image generation.

To this end, we try to use the Appearance code of image as the identity specific component and structure code as identity independent component. Thus based on this assumption, we can say that any other image x_2 of same modality as x_1 and same identity y_1 will have appearance code a_2 which will be similar to a_1. Thus keeping the structure code s same, we swap the appearance codes of x_1 and x_2 as shown in Figure 3a and reduce the l_1 loss between the original and reconstructed images.

$$L_{cross} = \mathbb{E} \left\| x_1 - G(a_2, s_1) \right\|_1$$

This equation is justified by the fact that s_1 will encode structural properties of x_1 and a_2 will have identity properties of x_1 so the output should look similar to x_1. We include this loss for image pairs x_1, x_2 to be all possible rgb-rgb pairs and ir-ir pairs. Note that we cannot include rgb-ir pairs having same identity over here as we do not have paired images of rgb-ir having same pose in dataset also if we swap ir structure with rgb appearance we cannot get an ir image back, it would be an rgb image. All these other cases will get covered in Different identity image generation.

However this may also result in a trivial solution where entire information gets leaked in s since decoders tend to rely on features having more spatial information. To avoid this, firstly structure encoder is given just gray scale image so that it has to rely on appearance code for color information. Also to make sure that appearance code of same identities are close and for different identities it is far apart, identification losses are applied on the input images based on their appearance code as below.

$$L_{id\_orig} = \mathbb{E}[-log(p(y_1|x_1))]$$

where $p(y_1|x_1)$ is the probability that x_1 belongs to it's ground truth label y_1 based on it's appearance code. Note that here x_1 is the original input image and not the reconstruction after swapping.

Different identity image generation Given an image x_1 and identity y_1, having some modality either rgb or ir, we sample another image x_2 having different identity y_2 whose modality can be anything. Thus here we include all pairs rgb-rgb , ir-ir and rgb-ir. Again we extract the codes s_1, a_1 and s_2, a_2 for images x_1 and x_2 respectively. We again swap the appearance code of a_1 with that of a_2 and reconstruct the image. The superscript denotes the image providing the appearance and subscript denotes the image providing the structure code. Thus we have,

$$x_1^2 = G(a_2, s_1) \text{ and } x_2^1 = G(a_1, s_2)$$

However here we cannot use the pixel level supervision as both images are not of same identity. Thus here we use the latent code reconstruction loss such that we can again get the structure and appearance codes from the generated image. [17]

$$L_{code}^s = \mathbb{E} \left\| s_1 - E_s(G(a_2, s_1)) \right\|_1 L_{code}^a = \mathbb{E} \left\| a_1 - E_a(G(a_1, s_2)) \right\|_1$$
$$L_{code} = L_{code}^s + L_{code}^a$$

Here also we apply the Identification loss on the reconstructed image based on it's appearance code. Note that here loss is applied on reconstructed image after swapping.

$$L_{id\_recon} = \mathbb{E}[-log(p(y_2|x_1^2))] \; L_{id} = L_{id\_orig} + \lambda_{id} L_{id\_recon}$$

Also Adversarial loss is added to ensure the distribution of generated data is as close as the real data distribution.

$$L_{adv} = \mathbb{E}[log(D(x_1)) + log(1 - D(G(a_1, s_2))]$$

3.4 Discriminative Module

The discriminative module is embedded in the generative module using the Appearance encoder as the backbone network. At the end of Appearance encoder, along with appearance code, we also get a vector $p \in \mathbb{R}^N$, where N is the number of classes. After swapping, the reconstructed image has structure code of one image and appearance code of another. Thus it is necessary to ensure that all information doesn't get leaked in one latent space. Hence a pre trained classifier trained on rgb and ir images is used to supervise this learning process. It is simply a baseline CNN trained with identification loss. To train the discriminative module, we minimise the KL divergence between the probability distribution $p(x_1^2)$ predicted by the discriminative module and the probability distribution $q(x_1^2)$ predicted by the pre trained classifier.

$$L_{dis} = \mathbb{E}\left[KL(p(x_1^2) \parallel q(x_1^2))\right]$$

We jointly train both the Encoders, Decoder and Discriminator to optimise the total loss. The total loss can be formulated as,

$$L_{total}(E_a, E_s, G, D) = \lambda_{recon}(L_{simple} + L_{cross}) + L_{id} + L_{code} + L_{adv} + L_{dis}$$

4 Training strategy

Generally, training a GAN is a tricky task and there can be training failure because of Mode collapse or convergence failure. In such cases it is difficult to make the Generator and Discriminator reach the equilibrium state. Thus we have evaluated our approach on two kinds of settings. In first setting we use the state of the art RGB-IR Reid network [13] as our baseline model. Then on top of this, we add our disentanglement model. Thus we used the output embeddings of this pre trained baseline model as the input to our Disentanglement model. In second setting, we removed the baseline model and our disentanglement model now directly gets the image as inputs. The reason for doing this two fold evaluation is that generally learning features in lower dimensional space is easier than high dimensional space. The higher the dimension, the more complex is the data distribution and the factors of variation are highly entangled making the disentanglement task more tedious. Thus we present our results using both these strategies and give inferences about which one is better. The following subsections elaborate the training strategy and network architectures used in both the cases.

4.1 Training in low dimensional embedding space

The baseline network [13] comprises of two main components: dual path network (one for RGB and one for IR) for feature extraction and bi-directional dual constrained top ranking loss for feature learning. We choose this network because it achieves a significant performance via end-to-end training. Note that since we do not have images coming as input, we cannot visualise whether the structure code is able to encode the structural information. So in this training, we simply call the two latent spaces as identity specific component s and identity independent component or modality component m.

The baseline model M_{bdtr} is first trained with bi-directional cross modality and intra modality top ranking loss and identity loss.[13] This network gives a 512 dimensional embedding x_{rgb} and x_{ir} for RGB and IR image respectively. We employ two Auto encoder networks, one for each modality. Thus we have two encoders E_{rgb}, E_{ir} and two decoders D_{rgb}, D_{ir}. Table 2 shows the architecture used for these networks. Each of the encoder gives two outputs, one is the identity specific component s and other is the identity independent component m. To enforce the constraints that s has identity specific information and m has identity independent information, we train the network by giving input as triplets of $[X_1, X_2, X_3]$

X_1 - RGB image of Id 1, X_2 - IR image having same Id as X_1, X_3 - IR image having different id

ENCODER-S	ENCODER-M	DECODER	CLASSIFIER
Input- 512	Input- 512	Input - S:64, M:64	Input- 512
1024, ReLU	1024, ReLU	512 ReLU : 512, ReLU	512, ReLU
2048, ReLU	2048, ReLU	2048 ReLU: 2048 ReLU	512, NUM_CLASS (395)
512, ReLU	512, ReLU	512 ReLU: 512 ReLU	
64. ReLU	64, ReLU	Output (S+M) - 512	

Table 2: Network Architecture for training in low dimension

$$m_{X1}, s_{X1} = E_{rgb}(X_1), \quad m_{X_2}, s_{X_2} = E_{ir}(x_2), \quad m_{X_3}, s_{X_3} = E_{ir}(X_3),$$

$$L_{simple} = \mathbb{E}\left\|X_1 - D_{rgb}(m_{X_1}, s_{X_1})\right\|_1 + \mathbb{E}\left\|X_2 - D_{ir}(m_{X_2}, s_{X_2})\right\|_1 + \mathbb{E}\left\|X_3 - D_{ir}(m_{X_3}, s_{X_3})\right\|_1$$

As discussed in section 3.3, the same identity reconstruction loss is applied by swapping the identity specific component of X_1 and X_2

$$L_{cross} = \mathbb{E}\left\|X_1 - D_{rgb}(m_{X_1}, s_{X_2})\right\|_1 + \mathbb{E}\left\|X_2 - D_{ir}(m_{X_2}, s_{X_1})\right\|_1$$

When different identity embeddings are swapped, we cannot apply l_1 loss as they are of different identity. So to ensure that the entire information is not leaked in any one of the latent spaces, we apply the identification loss on the reconstructed embedding by passing it through a pre trained classifier. This is similar to the Discriminative model described in section 3.4. Subscript denotes the embedding giving the m part and superscript denotes the embedding giving the s part.

$$X_1^3 = D_{rgb}(m_{X_1}, s_{X_3}), X_3^1 = D_{ir}(m_{X_3}, s_{X_1})$$
$$L_{dis} = \mathbb{E}\left[-log(p(y_3|X_1^3))\right] + \mathbb{E}\left[-log(p(y_1|X_3^1))\right]$$

where $p(y_3|X_1^3)$ is the probability that X_1^3 belongs to identity id y_3 and similarly for other term.

We also add a Triplet loss with X_1 as the anchor, X_2 as positive and X_3 as negative.

$$L_{triplet} = \max(0, m + \left\|s_{X_1} - s_{X_2}\right\|_1 - \left\|s_{X_1} - s_{X_3}\right\|_1)$$
$$L_{total} = L_{simple} + L_{cross} + L_{dis} + L_{triplet}$$

Kindly note that there are many more losses that can be applied here, like the latent code reconstruction loss, the identity loss on identity specific component s and also Adversarial loss on the reconstructed embedding. However simply adding all losses which looks sensible may not always help the training. Thus by careful observations, we have found out which loss may help in making s part identity specific, only those losses we have added in this Embedding space training. Table 6 shows the mAP and cmc values for different loss configurations.

Observations We can see some drawbacks with this approach on training in embedding space. First of all, although it's easier to train the network, we cannot visualise the latent dimensions. Thus it is difficult to evaluate the disentanglement performance. Secondly, separation into s and m will make sense only if the embedding which we get as input will itself have an s and m part, which is again difficult to predict as we cannot visualise the embedding. Thirdly, the mAP which we get at the final layer can be better than the mAP obtained in [13] only if we can claim there is some redundant information in the embedding which we try to remove. Making this claim again is not feasible, thus the mAP in [13] acts as a max limit to which we can take our mAP. Owing to these reasons, we then move to the High dimensional image space analysis and evaluate our disentanglement approach.

4.2 Training in high dimensional image space

The losses used here are same as described in section 3. Note that here we do not use separate encoder and decoder for different modalities. First reason is, if we use separate models, it will make the network huge and in turn difficult to train the model. So we started with single network for both modalities. Also this will make the network encode features common in both RGB and IR. The architecture details is given in Tables for E_s (Table 3), G (Table 4) and D (Table 5). All the network architecture is followed as done in [15]. E_a is based on Resnet50 [4] model and we remove the Global average pooling layer and FC layer and append it with adaptive max pooling layer. E_s has

Layer	Parameters	Output Size
Input	-	1 x 256 x 128
Conv1	[3 x 3, 16]	16 x 128 x 64
Conv2	[3 x 3, 32]	32 x 128 x 64
Conv3	[3 x 3, 32]	32 x 128 x 64
Conv4	[3 x 3, 64]	64 x 64 x 32
ResBlocks	$\begin{bmatrix} 3X3, 64 \\ 3X3, 64 \end{bmatrix}$ x 4	64 x 64 x 32
ASPP	[1 x 1, 32] $\begin{bmatrix} 3X3, 64 \\ 3X3, 64 \end{bmatrix}$ x 4	64 x 64 x 32
Conv5	[1 x 1, 128]	128 x 64 x 32

Table 3: Architecture of Structure Encoder E_s

Layer	Parameters	Output Size
Input	-	128 x 64 x 32
ResBlocks	$\begin{bmatrix} 3X3, 128 \\ 3X3, 128 \end{bmatrix}$ x 4	128 x 64 x 32
Upsample	-	128 x 128 x 64
Conv1	[5 x 5, 64]	64 x 128 x 64
Upsample	-	64 x x256 x 128
Conv2	[5 x 5, 32]	32 x 256 x 128
Conv3	[3 x 3, 32]	32 x 256 x 128
Conv4	[3 x 3, 32]	32 x 256 x 128
Conv5	[1 x 1, 3]	3 x 256 x 128

Table 4: Architecture of Generator G

Layer	Parameters	Output Size
Input	-	3 x 256 x 128
Conv1	[1 x 1, 32]	32 x 256 x 128
Conv2	[3 x 3, 32]	32 x 256 x 128
Conv3	[3 x 3, 32]	32 x 128 x 64
Conv4	[3 x 3, 32]	32 x 128 x 64
Conv5	[3 x 3, 64]	64 x 64 x 32
ResBlocks	$\begin{bmatrix} 3X3, 64 \\ 3X3, 64 \end{bmatrix}$ x 4	64 x 64 x 32
Conv6	[1 x 1, 1]	1 x 64 x 32

Table 5: Architecture of Discriminator D

Methods	Indoor Search	
	r1	mAP
$L_{simple} + L_{dis}$	1.33	9.07
$L_{simple} + L_{dis} + L_{cross}$	8.44	20.43
$L_{simple} + L_{dis} + L_{cross} +$ $L_{triplet}$ (random triplet)	14.25	28.85
$L_{simple} + L_{dis} + L_{cross} +$ $L_{triplet}$ (hard triplet)	**18.20**	**34.56**

Table 6: Results on SYSU-MM01 dataset with training on extracted embedding

an Atrous spatial pyramid pooling (ASPP) layer [2] which contains dilated convolutions and can be used to exploit multi scale features. The training strategy used also follows [15]. The hyperparameter λ_{id} is set as 0.5, as initially the reconstructed images are not good and keeping higher value may make the training unstable. As per the common practice in literature of image to image translation, [16][7][5] λ_{recon} is set as 5. L_{dis} , L_{code} are added only after generation quality is stable aroud after 30K iterations.

Figure 4: Top 10 retrieved images.

4.3 Experimental settings

Dataset Experiments were conducted on the publicly available large scale VI-Reid dataset, SYSU-MM01 [12]. The dataset contains images captured by 6 cameras, including two IR cameras and four RGB cameras. This dataset is challenging since some images are captured in indoor environment while some in outdoor environment. It contains 491 identities with fixed 296, 99, 96 identities in training, validation and testing set respectively. We use both training and validation data for training which involves 395 identities having total 22,258 RGB images and 11,909 IR images. The test set has

total 3803 IR images which are used as Query and randomly selected 301 RGB images are chosen as Gallery set.

Evaluation protocol Out of four RGB cameras, two are indoor(cam1, cam2) and two are outdoor(cam4, cam5). Out of two IR cameras, one is indoor(cam3) and one is outdoor(cam6). There are two modes for evaluation, Indoor-search mode and All-search mode. In Indoor-search, RGB cameras 1,2 (excluding camera 4,5) are used as Gallery set and IR camera 3 (excluding camera 6) is used as Query set. In All-search mode, all RGB cameras are used as Gallery set and all IR cameras are used as Query set. Indoor search is less challenging compared to All-Search. For each of the modes, we adopt Single shot and Multi shot search settings. In Single shot, we include one image per camera per identity, and in Multi shot, ten images per identity in Gallery set. For Query set, all IR images are included in both settings. For evaluating the performance, we use Cumulative Matching characteristic (CMC) [10] and Mean average Precision (mAP). We repeat the above evaluation ten times, with random split of Gallery set and report the average performance.

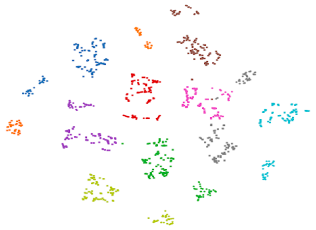


Figure 5: T-SNE plot for 10 random identities. Square denotes RGB data points and circle denotes IR data points.

5 Results and Conclusion

Table 7 and table 8 show the mAP and cmc values for All search and Indoor search mode respectively. Also comparison with the SOTA methods is done. It can be seen that our method gives comparable results to the existing methods which proves our claim that disentangling identity information can help in person Reid. Also Figure 4 shows the T-SNE [8] plot for ten random identities based on their appearance codes, which shows that the code for same identity images of both RGB and IR have come closer in the Appearance space. Also 5 shows the top 10 retrieved images for an IR image.

Methods	All Search							
	Single-shot				Multi-shot			
	r1	r10	r20	mAP	r12	r10	r20	mAP
BDTR	17.01	55.43	71.96	19.66	-	-	-	-
SDL	28.12	70.23	83.67	29.01	-	-	-	-
cmPIG	38.1	80.7	89.9	36.9	45.1	85.7	93.8	29.5
Hi-CMD	34.94	77.58	-	35.94	-	-	-	-
Method1	45.44	88.85	96.37	46.84	47.73	90.30	96.48	40.12
Method2	**52.51**	**92.01**	**97.05**	**49.51**	**56.14**	**93.56**	**98.34**	**43.81**

Table 7: Results on SYSU-MM01 dataset with All Search mode, Method1 : Training without pre-trained network, Method2: Training on model pretrained on Market dataset

Methods	Indoor Search							
	Single-shot				Multi-shot			
	r1	r10	r20	mAP	r12	r10	r20	mAP
SDL	32.56	80.45	90.67	39.56	-	-	-	-
cmPIG	43.8	86.2	94.2	52.9	52.7	91.1	96.4	42.7
Method1	51.95	95.24	98.78	63.11	52.65	94.10	98.35	51.22
Method2	**53.53**	**93.25**	**97.74**	**63.12**	**62.16**	**95.61**	**98.77**	**55.46**

Table 8: Results on SYSU-MM01 dataset with Indoor Search mode, Method1 : Training without pre-trained network, Method2: Training on model pretrained on Market dataset

References

[1] Yoshua Bengio. Deep learning of representations: Looking forward. In *International Conference on Statistical Language and Speech Processing*, pages 1–37. Springer, 2013.

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[3] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, volume 1, page 2, 2018.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[6] Kajal Kansal, AV Subramanyam, Zheng Wang, and Shin'ichi Satoh. Sdl: Spectrum-disentangled representation learning for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[7] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018.

[8] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[9] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in neural information processing systems*, pages 5040–5048, 2016.

[10] Hyeonjoon Moon and P Jonathon Phillips. Computational and performance aspects of pca-based face-recognition algorithms. *Perception*, 30(3):303–321, 2001.

[11] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 618–626, 2019.

[12] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017.

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

[14] L. Zheng, Y. Yang, and Q. Tian. Sift meets cnn: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1224–1244, 2018.

[15] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2138–2147, 2019.

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

[17] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 2017.