Unsupervised Cross Modality Person Re-Identification

by

Sagar Gupta

A thesis submitted in partial fulfillment for the
degree of Master of Technology

under supervision of
Dr. A.V. Subramanyam
Department of Electronics and Communication Engineering
Indraprastha Institute of Information Technology, Delhi

July 2020

Certificate

This is to certify that the thesis titled "Unsupervised Cross Modality Person Re-Identification" being submitted by Sagar Gupta (Roll No. MT18174) to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original work carried out by him under my supervision. In my opinion, thesis has reached the standards fulfilling the requirements of the regulations relating to the degree. The results contained in the thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

Date: _____

Dr. A.V. Subramanyam
Associate Professor
Department of Electronics and Communication Engineering
Indraprastha Institute of Information Technology, Delhi
New Delhi 110020

Abstract

Unsupervised Person Re-Identification (Re-ID) suffers severely from the gap in the modality. Many factors pose a challenge to the task, including occlusions, lightning conditions, pose changes, among several others. Various works try to use different methods to address the issue while we tried to solve it using GANs. We created images of another domain, conserving the identity of the person while changing the modality. It may so happen that a person moves from a well-lit area to an area where the light is way too low to be detected by the visual sensors. In such a case, the camera switches to IR, and the camera gets images in the Infrared spectrum. The method adopted for the generation of images is cycleGAN combined with pose loss and identity loss which further comprises of style loss and content loss. We are intent on getting IR images of a person with different pose whose RGB photos we have while preserving the identity. Besides, we aim to apply the existing state of the art techniques for Unsupervised Person Re-Identification for gauging our images.

Acknowledgements

I would like to express my sincere gratitude to Dr. A.V. Subramanyam for giving me the opportunity to work on my thesis under her guidance. His advice, expertise and experience have been invaluable for my entire research work as well as my career.

I am grateful to my parents for supporting me with love and guidance for my career.

Contents

Certificate **i**

Abstract **ii**

Acknowledgements **iii**

List of Figures **v**

List of Tables **vi**

List of Symbols **vii**

1 Introduction **1**
1.0.1 Structure of the Work 2

2 Related Works **3**
2.0.1 Cross Modality 3
2.0.2 Cyclegan 4
2.0.3 Poseloss 5
2.0.4 Identity Loss 5
2.0.5 Unsupervised Person Re-Identification 6

3 Proposed Approach **7**

4 Experiments **11**

5 Experiment Results and comparisons **16**

6 Conclusion and future scope **20**

Bibliography **21**

List of Figures

3.1 X,Y are two domains. In our case they are IR and RGB. G_A and G_B are the two generators. G_A and G_B can be thought of as a mapping function from X to Y and vice-versa. (b) Applying the G_A and then G_B will give us back a $\hat{X}$. The difference between X and $\hat{X}$ will give us the cycle consistency loss. Similarly in (c) the G_B and G_A are swapped as well as the X and Y are swapped 7

3.2 The generators G_A which is for RGB to IR conversion and G_B which is for IR to RGB conversion. Concatenation of the original image $I_{\mathbf{po}}$ and the target pose $\mathbf{P}f$ is sent to the G_A. The generator G_A converts it to an IR image with the fed pose. The new image is passed through the pose finding algorithm, and its pose is compared to the fed pose to calculate the poseloss. The new IR image $I_{\mathbf{pf}}$ is fed with the original pose $\mathbf{P}o$ to get back the original image. The cycleGAN loss ensures that the recovered and the original person are the same. The 3 images are fed into the vgg16 network to find the content and style loss. Only the output of the RGB-IR generator is shown whereas the IR-RGB generator follows a similar pattern but is not represented to make the diagram clearer. 8

3.3 Here we can see the marked 18 points joined together to show the pose of the person. It is a COCO based method. It is represented in the code as a .npy of 18*H*W where H is the height of the image and W is the width of the image. 9

3.4 In this image we get to know from which layers do we take the features for getting the style and content loss. 10

4.1 Initial results with lower number of epochs. (a) Real RGB Image, (b) Fake IR image, (c) Recovered RGB image , (d) Real IR image, (e) Fake RGB image (f) Recovered IR image 13

4.2 Intermediate results with lower number of epochs. (a) Real RGB Image, (b) Fake IR image, (c) Recovered RGB image , (d) Real IR image, (e) Fake RGB image (f) Recovered IR image 14

4.3 Final results. (a) Real grayscale Image, (b) Fake IR image, (c) Recovered grayscale image , (d) Real IR image, (e) Fake grayscale image (f)Recovered IR image 15

5.1 Comparison table with Unsupervised methods 18

5.2 Supervised Comparison table 19

List of Tables

5.1 Results with MMT [1] method 16
5.2 Results with OSNet[2] method 17
5.3 Results with labelled and unlabelled dataset with the method of ECN[3]. 17
5.4 Comparison table with normal triplet loss [4], SCE loss[5] and combination of both 18
5.5 Learning rate, epochs, weights and various other parameters were experimented with to achieve all the desired outputs. 19

List of Symbols

Symbol	Description
G_A	Generator A
G_B	Generator B
D_A	Generator A
G_B	Generator B
ψ	Vgg16 trained network
I_{po}	Image with the original pose
I_{pf}	Image with fake pose
$\mathbf{p}_o$	Original Pose
$\mathbf{p}_f$	Target Pose / Fake Pose
$\hat{I}_{po}$	Recovered Image with original pose

Chapter 1

Introduction

Person Re-Identification (PRID) is one of the most critical tasks of computer vision. It aims to match people's images over different overlapping cameras. Such a system suffers from various issues such as changes in the person's view, posture, lightening condition, background. Various methods have been proposed for the same including distance metric learning [6],[7],[8] and feature learning [9],[10]

However, a few papers have worked on cross-modality person re-identification [11],[12]. Cross modality is the term when the Re-ID is done between 2 different domain. It can be Thermal and RGB or Thermal and UV. In our case, it is RGB and IR domains. In many of the applications, the camera may work in a dual-mode often combining RGB-IR and thermal cameras. Modern surveillance cameras include automatic switching modes when an insufficient amount of light is present for RGB to sustain. The key challenge is the unavailability of information to draw parallels between IR and RGB images. Learning a robust metric is also quite difficult.

Recently SYSU dataset was released by [12] comprising of frames of 6 cameras. With camera number 3 and 6 having IR images and the rest of the cameras having standard RGB images, it has one of the most comprehensive datasets with subjects in both IR and RGB. It facilitates researchers to find tools and methods for a better system to work in both conditions.

We face two challenges. We need to develop a system to convert RGB images into IR images with reliable accuracy to produce good quality images. Secondly, we try to replicate a method of unsupervised learning to create good mAP and Rank-1, rank-5, rank-10 Rank 20 accuracies. The work is organized in the following manner. In the following section, we present the related works in the field of both cross-modality and unsupervised person re-Identification. Chapter 3 gives the proposed approach and

the theory behind it. The next section talks about the experimental setup and the experiments performed. Subsequently, chapter 5 elaborates the results and loss graphs and compares the images with state of the art unsupervised person re-identification techniques. We finally summarize our conclusion in chapter 6.

1.0.1 Structure of the Work

This thesis is divided in 6 chapters.

- **Chapter 2: Related Works**
  This chapter covers all the existing methods to bridge the gap between IR and RGB domains and also talk about the state of the art unsupervised person re-id techniques.
- **Chapter 3: Proposed Approach**
  Proposed approach identifies the methods and explains the various losses taken to optimize the algorithm.
- **Chapter 4: Experiments**
  This section tells about the various experiments performed and the created images.
- **Chapter 5: Experimental Results**
  In this chapter, results and analysis is done on the various outcomes. Various loss graphs are plotted along with the resulting images.
- **Chapter 6: Conclusion**
  In this chapter, we conclude the research done.

Chapter 2

Related Works

The related works section is further broken down according to the topics

2.0.1 Cross Modality

Various papers of person re-identification and cross-modality were referenced. Some papers of facial re-identification were also consulted to find the best strategy for cross-modality.

[11] has a DCNN generator to represent modality invariant features and a modality (RGB-IR) discriminator to differentiate between RGB and IR. The generator used ID loss and cross-modality triplet loss.

In [13] Semantic labels extracted from a face parsing network are used for computing a semantic loss function. Semantic cues(features) extracted around eyes, nose, mouth is considered to be beneficial for synthesizing identity-preserved face images. Larger weights are assigned to features extracted more deeply. Semantic-guided generative adversarial network (SGGAN) used which was trained on PCSO dataset(having both thermal and RGB images). Thus learning the relationship between the two domains and was then fine-tuned on ARL dataset. Face matching algorithms like AM-softmax and MobileFaceNet were used as a baseline for the method.

[14] is based on pix2pix and DRGAN. It aims to create fake images in the visible domain and run a normal facial recognition software. [15] synthesizes more photo-realistic images. The research claims that the polarimetric images have more geometric features. Simple GAN with losses used is for getting the translated images. GAN subnetworks are used for guidance and getting semantic features. It is used while training to make the images more photo-realistic.

The authors in [16] aim to map heterogeneous multimedia into common hamming space. In the proposed model, when given the data of a modality, the generator tries to fit distribution over the manifold structure. The authors in [17] try translating VIS face images into fake Near Infrared(NIR) images whose distribution are intended to approximate those of true NIR images.

[16] have a given data of a modality, and the generative model tries to fit the distribution over the manifold structure. It also selects informative data of the other modality and includes it in a way while fitting the distribution over the manifold structure of the data being fitted. This is done for making a fool of the discriminator. The discriminator tries to find out which data is from which distribution. Correlation graph is plotted among the manifold structure of the different modalities, so data between different modality but same manifold can have smaller hamming distance and promote retrieval accuracy. It is integrated into the GANs network. Authors generate hash codes pairwise of the data in each modality. The generator tries to generate pairs from the data, while discriminator can make the correlation matrix between the data and create manifold pairs. Discriminator discriminates between manifold pairs and generated pairs.

[18] transfers the style of the images from the source domain to the target domain. Label smooth Regularization(LSR) is employed where soft labels are used. Vanilla version is without LSR while the full version is with LSR. In the vanilla version, we are learning image to image translation between two image pairs. After the translation, we use the style transferred images to train the CNN. Labels are borrowed directly from the photos from whom the new images are generated. For each pair of a camera with different views(styles), an image to image translation model is learnt.

The aim of the authors in [19] is to generate fake IR images based on real RGB images via a pixel alignment module, and then match the generated fake IR images and authentic IR images via a feature alignment module. In [20] the authors tackle the above limitation by proposing a novel cross-modality shared-specific feature transfer algorithm (termed cm-SSFT) to explore the potential of both the modality-shared information and the modality-specific characteristics to boost the re-identification performance. They model the affinities of different modality samples according to the shared features and then transfer both shared and specific features among and across modalities.

2.0.2 Cyclegan

[21] introduced the cyclegan to generate translations in the image. They experimented in a supervised manner and were able to give good results interchanging the images of one domain to another and vice-versa. Since paired data is not available from source

X domain to target domain, they tried to achieve a function G where G: X->Y and the distribution made by G(X) is indistinguishable from the Y. An inverse function F: Y->X is also introduced. This is achieved by using GANs [22]. A cycle consistency loss is introduced to enforce the F(G(X))==X. Wherever the paired data is unavailable, such a method can be used. The various applications included the season transfer, photo enhancement etc. We have utilized cyclegan in our framework for converting the images of the RGB images to IR and IR images to RGB.

2.0.3 Poseloss

Given a person's image and a pose, an image can be synthesized to produce a person with the given pose. [23] achieved it by proposing the novel Pose Guided Person Generation Network (PG 2) that allows synthesizing person images in arbitrary poses, based on an image of that person and a novel pose. Eighteen heatmaps are concatenated with the input image and given to the generator G1, which creates an image which is a coarse image. This coarse image is concatenated with the original image and fed to generator G2 to create a difference map. After the losses, the discriminator tries to differentiate between the (original + created coarse image) and (original + target image). This forces the generator 2 to make the coarse image as similar as possible to the original target image. The difference map is obtained in the second stage with G2 (the fully connected layer is removed in the second stage as they compress the information.) The uncompressed information from the (coarse image + input image) and the image from the previous stage is calculated to calculate the difference. Pose mask loss was shown to be showing better than the L1 results.

In [24] given an image and a target pose, the image is changed to have the target pose. This paper utilized nearest neighbourhood loss instead of the usual L1, L2 losses. The authors propose deformable skip connections to deal with this misalignment problem and "shuttle" local information from the encoder to the decoder driven by the specific pose difference. Joint transformation is done on each and every joint so as to convert each of the pose-joint. We have utilized PoseLoss [25] to get the pose of the generated images. It presents a real-time approach to detect the poses even in the IR domain using part affinity fields.

2.0.4 Identity Loss

[26] introduces a method to create artistic style images using Deep Neural Network (DNN). The authors utilized the representation of the photos using DNN to separate and recombine the style and content of the photos. We have utilized the style and

content loss to get good quality images after extracting and recombining the style and content of the IR and RGB images.

2.0.5 Unsupervised Person Re-Identification

The generated images are then tested based on various metrics to find the efficacy of the system. Mutual Mean Teaching (MMT) introduced in [1] focusses on refining of labels for domain adaptation. It utilized noisy pseudo labels to improve feature representation on the target domain in an online peer teaching manner. Novel soft softmax triplet loss was used along with classification loss for achieving acceptable results. Pseudo labels are generated by clustering. The authors in [2] designed a residual block made up of many convolution streams with each of them finding representations at a fixed scale. Multi-scale features are fused with input dependent channel-wise weights. The method captures various scales and their various combinations.

Authors in [4] try to optimize the embedding space such that the data points belonging to the same identity are brought closer while the ones belonging to different identities are kept farther. Hard triplet was used in the process, which refers to taking the farthest positive pair and the closest negative pair. The anchor point and the positive point (which belongs to the same class) are closer than the negative data point(belonging to a different class).

Conventional methods are mainly to reduce feature distribution gap between the source and target domains. [3] takes into account the intradomain variations of the target domain. They have tried to generalize the invariance into three types:- exemplar-invariance, camera-invariance and neighbourhood-invariance. So i.e. it takes two datasets, their images and tries to reduce the intra-identity distance and increase the inter identity distance.

[5] gives a new approach to learning with noisy labels. For the learning of hard classes, the cross-entropy term has to become more tolerant of noisy labels. Based on KL divergence they proposed Symmetric Cross-Entropy Learning algorithm which takes into account the reverse cross-entropy

Chapter 3

Proposed Approach

The proposed approach uses GANs in a cyclegan manner. GANs were introduced in [22]. It was introduced as a min-max game played between the discriminator and generator to one-up each other. This keeps going until finally, the discriminator is unable to differentiate between the image created by the generator and the real image.

The original cyclegan takes an image converts it and then reconverts it back to the original image. This is known as cycle loss. In our case, the two generators are used for conversion from RGB to IR and from IR back to RGB 3.1. Concatenation of image and pose is given as input, whereas the output theoretically is the same person in the IR domain with the pose that was given as the input. Recovering with the original pose gives the original person back in the RGB domain with the given pose as shown in figure 3.2.

In the 3.2 figure we can see two generators G_A and G_B. The first generator is for converting the RGB to IR images. The second generator is for converting the IR to RGB images. The target pose $I_{\mathbf{pf}}$ is concatenated and fed to the generator to translate

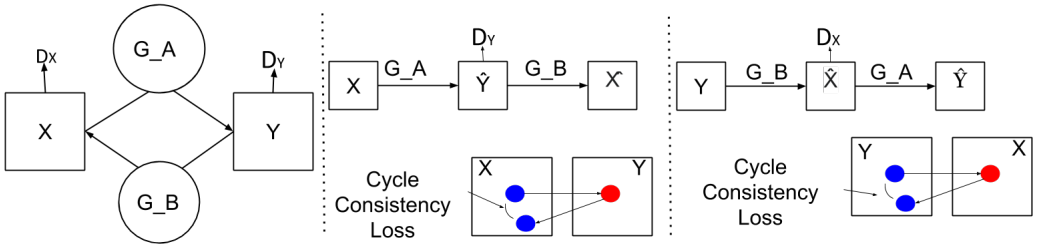


FIGURE 3.1: X,Y are two domains. In our case they are IR and RGB. G_A and G_B are the two generators. G_A and G_B can be thought of as a mapping function from X to Y and vice-versa. (b) Applying the G_A and then G_B will give us back a $\hat{\text{X}}$. The difference between X and $\hat{\text{X}}$ will give us the cycle consistency loss. Similarly in (c) the G_B and G_A are swapped as well as the X and Y are swapped

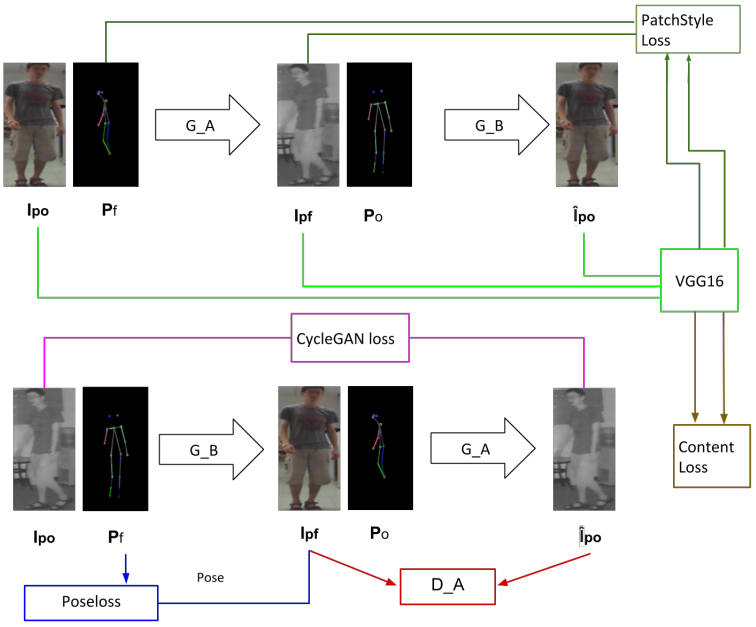


FIGURE 3.2: The generators G_A which is for RGB to IR conversion and G_B which is for IR to RGB conversion. Concatenation of the original image $I_{\mathbf{po}}$ and the target pose $\mathbf{P}f$ is sent to the G_A. The generator G_A converts it to an IR image with the fed pose. The new image is passed through the pose finding algorithm, and its pose is compared to the fed pose to calculate the poseloss. The new IR image $I_{\mathbf{pf}}$ is fed with the original pose $\mathbf{P}o$ to get back the original image. The cycleGAN loss ensures that the recovered and the original person are the same. The 3 images are fed into the vgg16 network to find the content and style loss. Only the output of the RGB-IR generator is shown whereas the IR-RGB generator follows a similar pattern but is not represented to make the diagram clearer.

it into an image with the similar image with the transformed into IR domain with the fed pose. This obtained image is fed with the original pose $I_{\mathbf{po}}$ to get the original image back in the RGB domain.

A pose finding algorithm [25] with the ability to find the poses in both RGB and IR was to used to calculate the poseloss and the identity losses later. The pose is represented by its joints as shown in figure 3.3

Content loss and style loss was originally coined in the [26] and later in the paper of Image Style Transfer Using Convolutional Neural Networks [27], neural style transfer has found its place into many of the papers. We have used the loss to preserve the identity of the images and change the poses of the same. The identity loss is represented as the sum of the patch-style loss and the content loss represented in eq. 3.1

$$\mathcal{L}_{Id} = \mathcal{L}_{Content}(\Psi, I_{po}, \hat{I}_{po}) + \lambda \mathcal{L}_{Patch\_Style}(\Psi, I_{po}, I_{pf}, \mathbf{p}_o, \mathbf{p}_f) \tag{3.1}$$

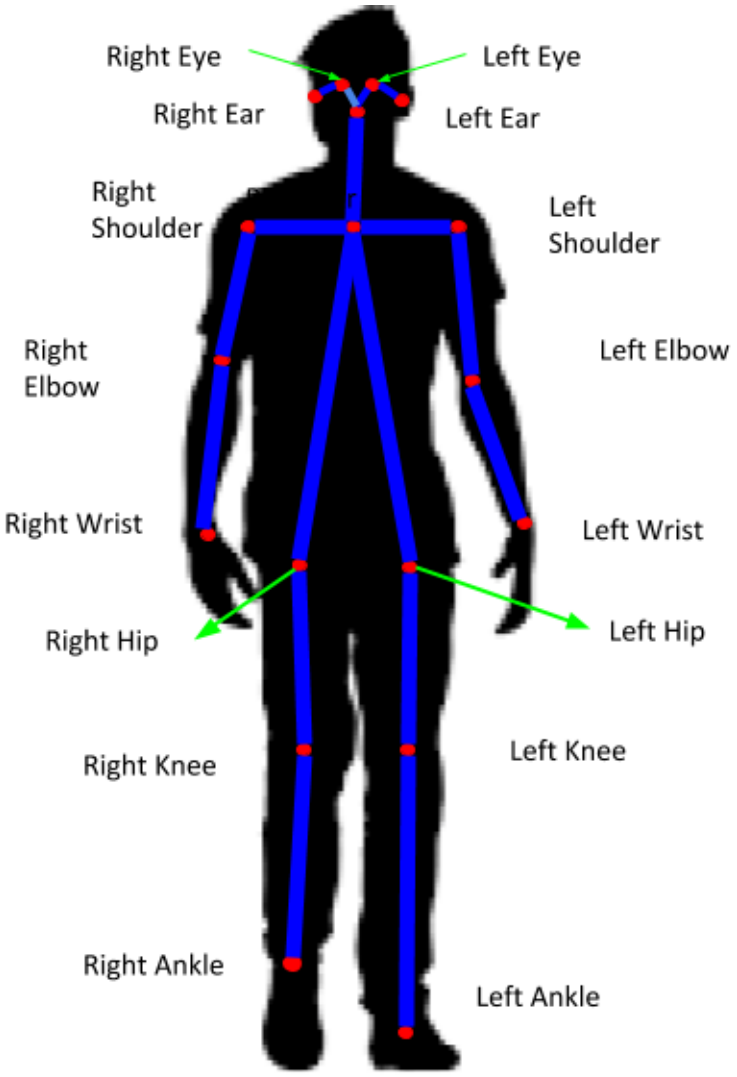


FIGURE 3.3: Here we can see the marked 18 points joined together to show the pose of the person. It is a COCO based method. It is represented in the code as a .npy of 18*H*W where H is the height of the image and W is the width of the image.

Here in eq. 3.1 Ψ represent the pre-trained VGG16. I_{po} represents the image with the original pose, whereas the I_{pf} represent the generated image. The p_o represents the original pose whereas the p_f represent the the the fake pose that has been fed. The lambda is the weight assigned to the patch style vs content loss.

$$L_{Content} = ||(\Psi_z(I_{po}) - \Psi_z(\hat{I}_{po})||_2^2 \tag{3.2}$$

The content loss in eq. 3.2 is the L2 difference between the VGG16 features of the image and the recovered image.

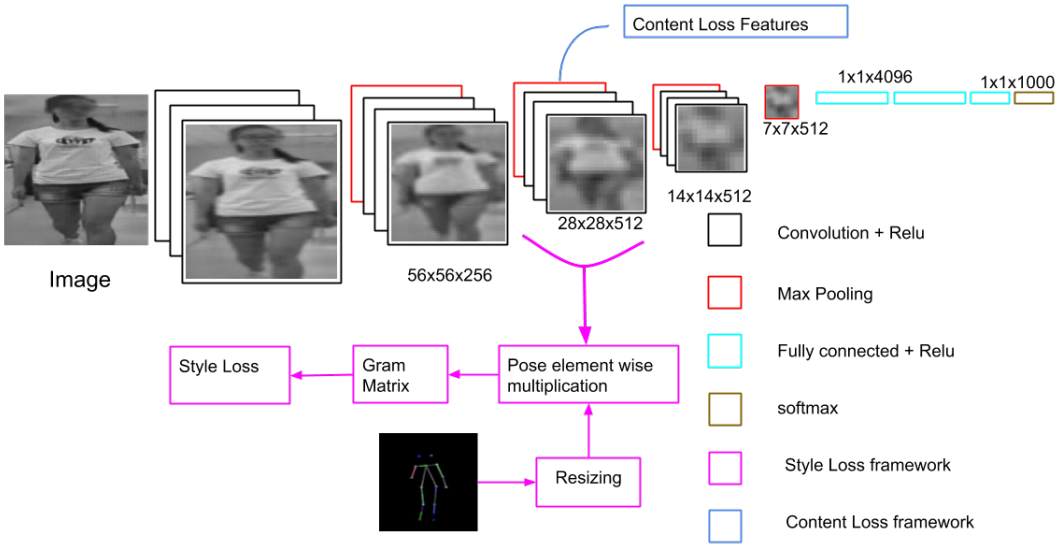


FIGURE 3.4: In this image we get to know from which layers do we take the features for getting the style and content loss.

$$\mathbf{X}_{po}, i = \mathbf{B}_{po}, i \ . \ \Psi_z(\mathbf{I}_{po}) \qquad \forall i \in \{1, ...N\} \tag{3.3}$$

$$\mathcal{L}_{Patch\_Style} = \frac{1}{N} . \sum_{i}^{N} \left(\frac{G_{po}, i - G_{pf}, i}{H'.W'} \right)^2 \tag{3.4}$$

$\mathbf{B}_{po}, i$ is the downsampled probability map of the associated to the pose $\mathbf{p}_o$. Representation of a patch style is then captured by the correlation between the different channels of its hidden representations $\mathbf{X}_{po}, i$ using the spatial extend of the feature maps as the expectation. Gram matrix $\mathbf{G}_{po}, i$ is found for each of the patch and mse is found between gram matrix of same joints in both images I_{po} and I_{pf} depicted in equation 3.4. Representation of the image is given in image 3.4

$$\begin{aligned} \mathcal{L} =& \mathcal{L}_1(G, D_i, I_{po}, \mathbf{p}_f) + \lambda_P \mathcal{L}_P(G, \Phi, I_{po}, \mathbf{p}_f) \\ &+ \mathcal{L}_1(G, D_i, I_{pf}, \mathbf{p}_o) + \lambda_P \mathcal{L}_P(G, \Phi, I_{pf}, \mathbf{p}_o) \\ &+ \lambda_{Id} \mathcal{L}_{Id} \end{aligned} \tag{3.5}$$

Linear combination of all the losses is taken as the final loss and back propagated to the generators as shown in the equation 3.5 . The value weights are decided by the ablation study in table 5.5.

Chapter 4

Experiments

Basic cycleGAN was used as the template for the code. RGB images and a target pose were fed to the generator to produce the IR images with the given pose. Another generator was fed with this IR image and the original pose to get the original image with the original pose back. In a parallel setting, IR images were first fed along with the target pose to the IR generator to produce an RGB image with the changed pose. Then the results along with the original pose, were fed to the IR generator to get the original IR image back. Discriminators were placed at the outputs of the generators, and the images were compared with each other as shown in figure 3.2.

For the purpose of detecting the poses, we checked various methods and compared the images to the stick figures. We checked the images with three different methods. Pose-gan pose estimator [28], Openpose pose estimator and Human pose convolution machines [29]. The Pose-gan pose estimator worked very slowly and took a few seconds for each of the images to annotate. These pose annotators did not work on the IR images at the time. So we took a few days and annotated 18 points of L eye, R eye, L Ear, R Ear, Nose, Neck, L shoulder, R shoulder, L Elbow, R elbow, L wrist, R Wrist, L hip, R hip, L knee and R knee of around 400 images.

We kept on experimenting with all the available methods to find the poses of the IR images. It turned out that open pose couldn't work when the person was over the whole images and could be detected if he was in a part of the images. We padded the images on each side, and openpose started working perfectly with even the IR images. The openpose was made as a loss in the cyclegan and was fed to the generator. We observed that the cores kept getting dumped and the modified code was way slower than its individual code parts. Found out that the openpose kept accumulating the found poses and got the memory full. We decided to change the instance losses to batch losses in the newly created openpose.

The content and the style loss gets the content from one image and the style of another and merges them together. For content features a lower level convolution feature was taken after the image was fed into VGG19 as lower-level maintains the global arrangement of the image. For style loss, lower level outputs of 5 higher levels(z = 7) are taken where only the style of the image is present. We used these features as a loss with modifications. We multiplied the original image with the pose matrix of the target pose and took the content of the image that we had to convert. These two losses combined are named as identity two loss in the code. Ablation study was conducted to find out the weight of each of the losses 5.5. Earlier, the losses yielded different results when the online code was used. Even setting seed and using pre-trained weights made no difference, so we took a Keras pre-trained VGG and calculated the features ourselves. Using the framework of cyclegan and Keras for VGG lead to the complete exhaustion of the GPU and shutting down of the servers. The VGG was then changed to PyTorch, and the resources were freed up for more batch sizes leading to faster training reducing the time of training from 30 days to 12 -15 days.

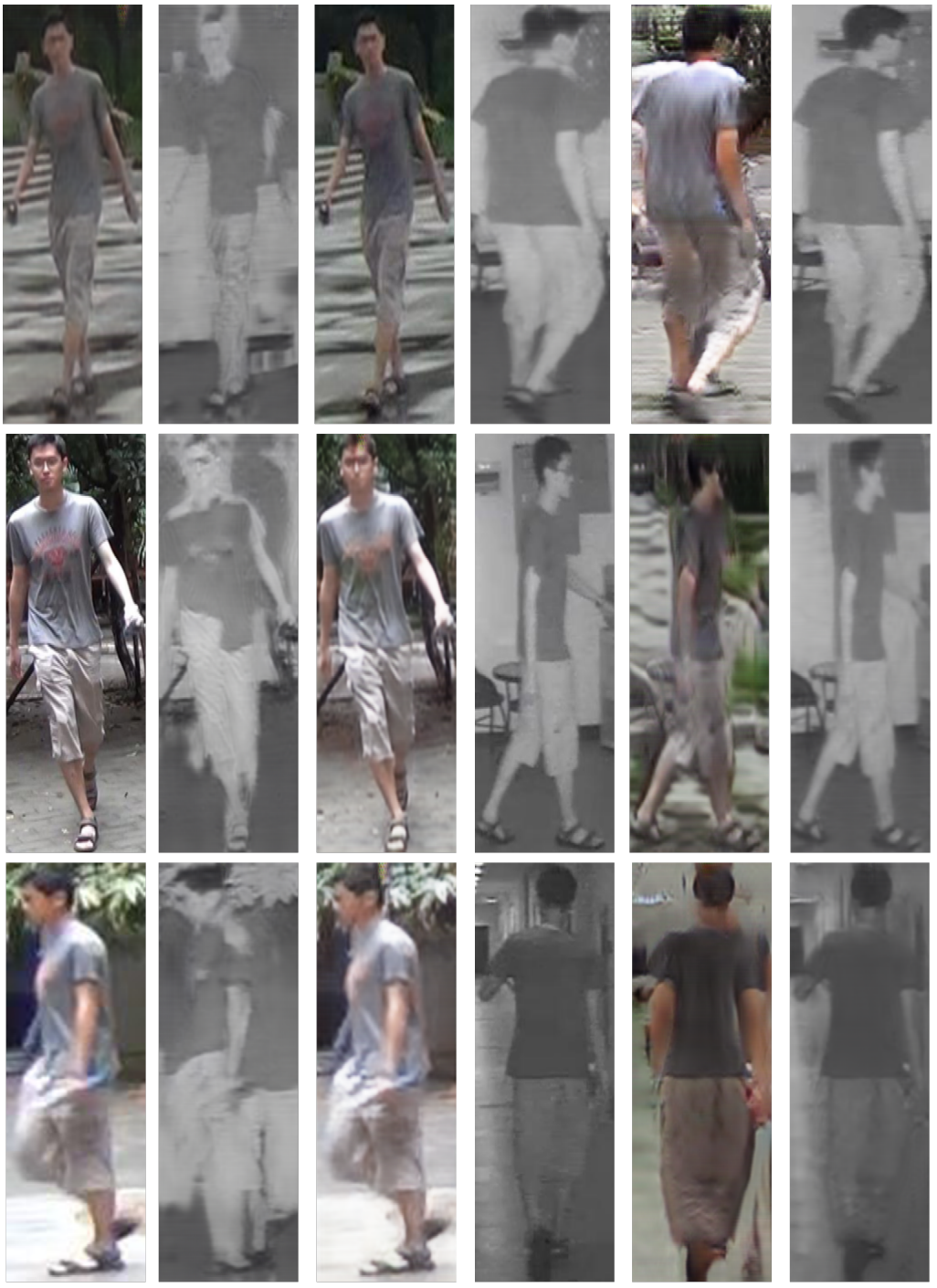
The pose loss and the Identity loss did not converge even after several permutations and combinations, whereas the IR images converted to the RGB and RGB converted to the IR.

In the initial results fig.4.1 the proper construction of the RGB image is happening but the IR images that are made are not upto the mark.

In the intermediate results fig.4.2 the IR and RGB images were getting constructed properly but the RGB images were getting randomly coloured once they were made from the IR images. To mitigate this all the RGB images were changed to grayscale.

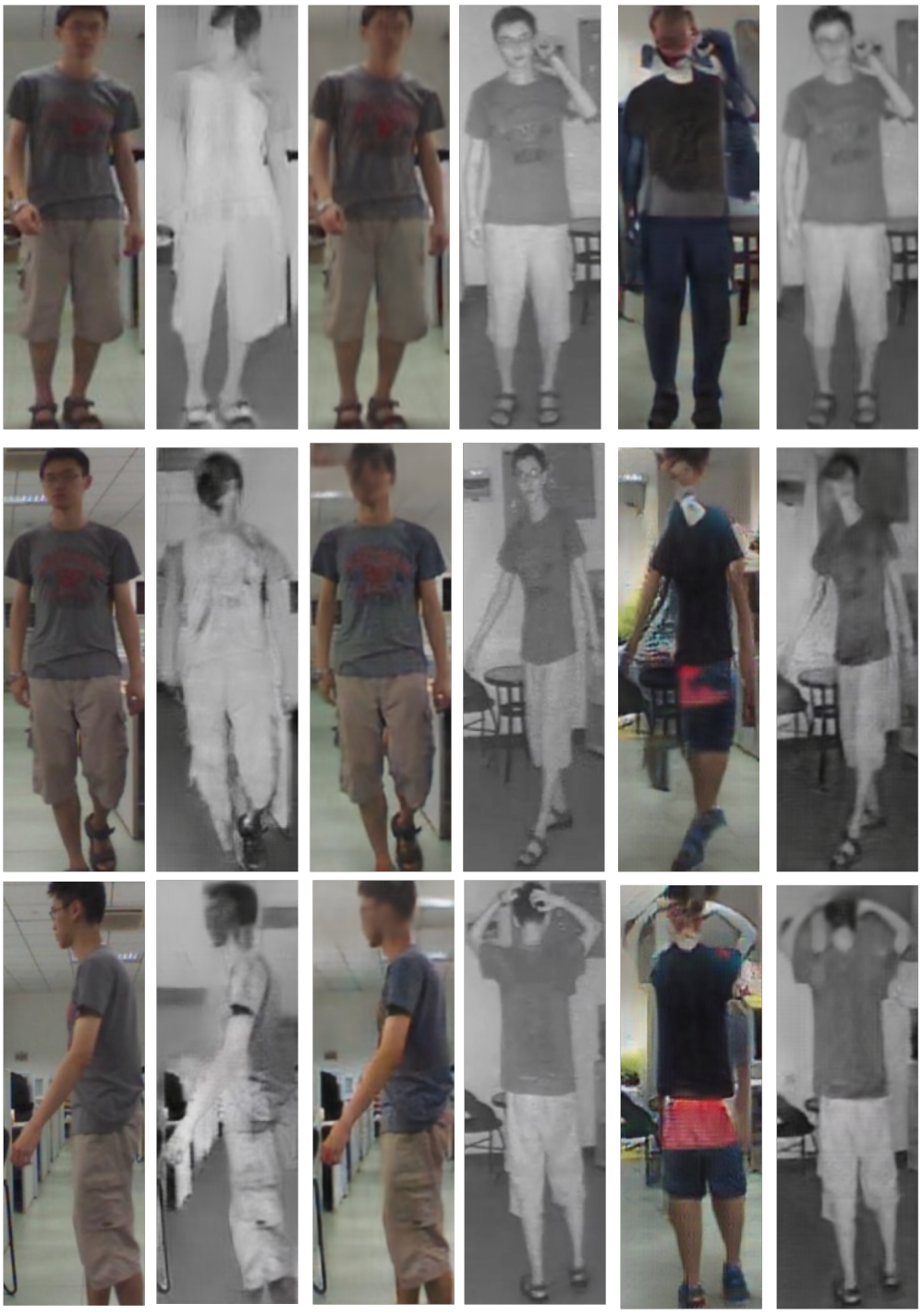
After the conversion to grayscale the image quality improved but in the case of the RGB to IR conversion the blurring occurs as show in fig.4.3.

As in MMT, using the cluster algorithms on the target domain, soft labels are assigned.



(a) (b) (c) (d) (e) (f)

FIGURE 4.1: Initial results with lower number of epochs. (a) Real RGB Image, (b) Fake IR image, (c) Recovered RGB image , (d) Real IR image, (e) Fake RGB image (f) Recovered IR image



(a) (b) (c) (d) (e) (f)

FIGURE 4.2: Intermediate results with lower number of epochs. (a) Real RGB Image, (b) Fake IR image, (c) Recovered RGB image , (d) Real IR image, (e) Fake RGB image (f) Recovered IR image

(a) (b) (c) (d) (e) (f)

FIGURE 4.3: Final results. (a) Real grayscale Image, (b) Fake IR image, (c) Recovered grayscale image , (d) Real IR image, (e) Fake grayscale image (f)Recovered IR image

Chapter 5

Experiment Results and comparisons

This section talks about the various unsupervised algorithms.

In the Generator loss plot, the RGB-IR generator loss converged properly whereas the IR to RGB generator loss convergence faces some problem and its value oscillates after initially converging. When the pose did not change we looked at the discriminator loss. The discriminator loss converged way too fast to do any meaningful learning. The pose loss and the Identity loss did not converge, thereby preventing any pose change.

We applied GAN tricks of swapping the real and fake labels, reduced the learning rate by a tenth. The converging rate of the discriminator slows.

	mAP	**Rank1**	**Rank 5**	**Rank 10**
Market - 1501 and Sysu original (in percentage)	0.2	0	0.8	1.7
Sysu-Sysu results (in percentages)	0.2	0.2	1	1.7
Best reults Sysu- Syuresults (In percentage)	0.4	0.4	1.9	3.9

TABLE 5.1: Results with MMT [1] method

Various unsupervised algorithms were applied to the obtained images to test the efficacy of the system. Results were calculated with [1] and is given in table 5.1. The low accuracy is probably because of the reason that this algorithm makes clusters and the clusters of the IR and RGB images lie far from each other.

The next approach we applied is [2]. A residual block made up of multiple convolution streams is used for detecting feature at a certain scale. The results are indicated in table 5.2.

	mAP	Rank1	Rank 5	Rank 10	Rank20
market 1501	80.8	92.5			
Sysu original	2.1	0.8	1.6	3.3	5.26
Sysu cyclegan images	2.1	0.4	1.2	3.5	6.9
Sysu combined	1.7	0.3	0.6	3.2	3.6

TABLE 5.2: Results with OSNet[2] method

Next, we used [3], which uses labelled source domain and unlabelled target domain. For the labelled source domain we had the SYSU-MM01 dataset, for the unlabelled we used the cyclegan images that we produced. This method takes into account the intradomain variations of the target domain. The results are given in table 5.3.

CMC	R-1	R-5	R-10	R-20
9.2	6.9	15.1	19.2	24.1

TABLE 5.3: Results with labelled and unlabelled dataset with the method of ECN[3].

[5] proposes the use of noisy labels. Since in our case, the images generated does not have a hard label their label can be considered a noisy label of the image that they were created from. Taking the symmetric idea of KL divergence the authors have suggested the symmetric cross entropy which is the sum of the Reverse cross-entropy and cross-entropy. Various results with various combinations have been found. The Reverse Cross Entropy (RCE) loss was also combined with the triplet loss(with Batch hard strategy) and produced comparable results in the case of the Combined SYSU dataset. 4 types of datasets were used to find the comparable accuracies in the experiment. Sysu original (RGB + IR images), Sysu cyclegan (Grayscaled+IR images), Sysu combined (Sysu original+Sysu cyclegan), Sysu combined all cameras (using all cameras images and transformed images). The results have been indicated in table 5.4. The accuracies of the triplet loss combined with SCE lies very close to the ones without the triplet loss. Hence we can conclude that the triplet loss and hard labels are having almost no affect on the accuracies and the soft labels with SCE is giving good results.

Usage of the images to find the mAP and Rank accuracies of all the methods resulted in the comparison shown in figure 5.1.

Just as a proof of concept, the images were given hard labels and a supervised person re-Identification system was trained on the SYSU-MM01, SYSU cyclegan images and the combined SYSU dataset. We have applied the Beyond parts model of the paper [30] represented in the 5.2.

Dataset	Losses	mAP	top 1	top 2	top 5	top10	% of noisy labels
Sysu combined all cameras	**SCE**	0.2411	0.1793	0.1942	0.2335	0.2855	50
Sysu original images all cameras	**SCE**	0.2061	0.1461	0.1753	0.1927	0.2432	
Sysu combined all cameras	**Triplet + Noisy labels (scaled 0-1)**	0.248	0.1895	0.1925	0.2354	0.2829	50
Sysu combined all cameras	**Triplet**	0.1856	0.1492	0.1822	0.202	0.211	
Sysu original images all cameras	**Triplet + Noisy labels (scaled 0-1)**	0.1288	0.0964	0.099	0.1214	0.1423	10
Sysu original images all cameras	**Triplet**	0.1495	0.1147	0.1365	0.1462	0.167	
Sysu cyclegan images all cameras	**Triplet + Noisy labels (scaled 0-1)**	0.1358	0.0986	0.1203	0.1245	0.1528	50
Sysu cyclegan images all cameras	**Triplet**	0.1063	0.078	0.0997	0.1179	0.1356	

TABLE 5.4: Comparison table with normal triplet loss [4], SCE loss[5] and combination of both

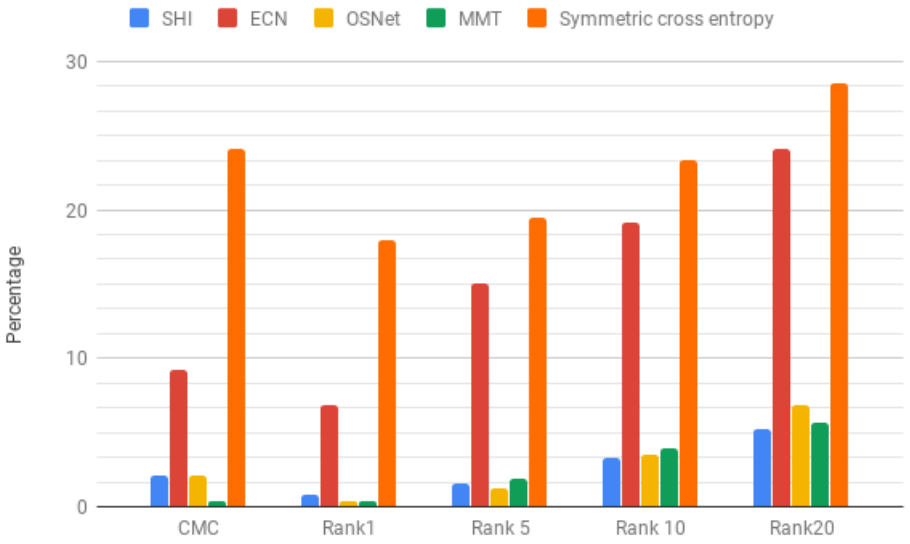


FIGURE 5.1: Comparison table with Unsupervised methods

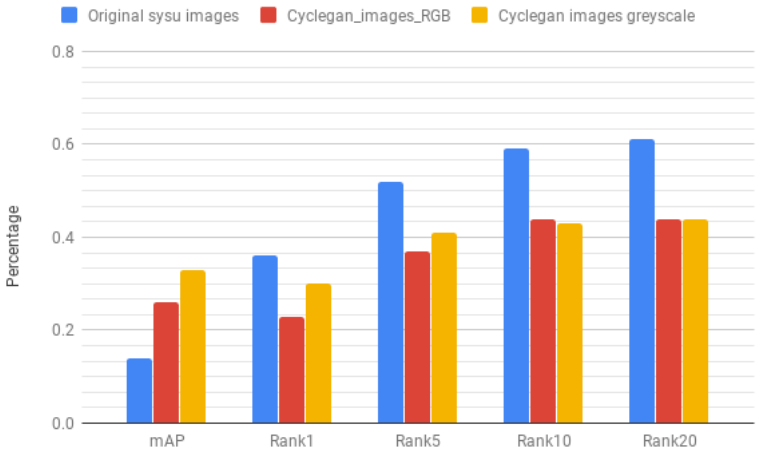


FIGURE 5.2: Supervised Comparison table

Iteration	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Learning Rate (e-4)	3	3	3	3		3	3	3	3	3	3	3	3	0.25	0.3	0.3	0.3	0.3
Epochs	100	100	100	150	200	300	300	300	300	300	300	300	300	300	300	300	300	300
Lambda Cycle	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
Lambda Pose	0	0	700	700	700	700	700	700	700	700	1.13	1.13	1.13	1.13	1.13	1.13	1.13	1.13
Lambda Identity (3e-3)	0	0	0	0	0	100	100	100	100	100	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Lambda Content	0	0	0	0	0	5	5	5	5	5	5	5	5	5	5	5	5	5
Lambda Style (e-5)	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Batch Size	4	4	4	4	4	2	1	1	1	4	8	8	8	8	8	8	8	8
Num Threads	4	4	4	4	4	8	8	8	8	8	4	4	4	4	4	4	4	4
Pose Changed	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Image Quality (improved or not)	X	X	X	✓	✓	✓	X	X	X	X	✓	✓	X	✓	X	X	X	X
Generator, Discriminator loss (converged or not)	X	X	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Pose Loss (converged or not)	X	X	X	✓	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Cyclegan Loss (converged or not)	X	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Identity Loss (converged or not)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

TABLE 5.5: Learning rate, epochs, weights and various other parameters were experimented with to achieve all the desired outputs.

Chapter 6

Conclusion and future scope

We have tried to obtain good quality images using the pose-change GANs. We have applied Style loss and Content loss in addition to the cyclegan loss in the training of both the generators and discriminators We were able to change the images from the IR to RGB and vice-versa. Despite our best efforts, we were not able to change the pose of the images. After getting the images, we combined them with the existing original SYSU dataset and used it for training various supervised and unsupervised Person Re-Identification methods. Method of symmetric cross-entropy was able to achieve comparable accuracies. The future scope may include changing the dataset to RegDB dataset may help. For the convergence of pose loss, a same domain dataset can be first used.

Bibliography

[1] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020.

[2] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3702–3712, 2019.

[3] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 598–607, 2019.

[4] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[5] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 322–330, 2019.

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

[7] Arulkumar Subramaniam, Moitreya Chatterjee, and Anurag Mittal. Deep neural networks with inexact matching for person re-identification. In *Advances in neural information processing systems*, pages 2667–2675, 2016.

[8] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by discriminative selection in video ranking. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2501–2514, 2016.

[9] 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.

[10] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1363–1372, 2016.

[11] 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.

[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] Cunjian Chen and Arun Ross. Matching thermal to visible face images using a semantic-guided generative adversarial network. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[14] Teng Zhang, Arnold Wiliem, Siqi Yang, and Brian Lovell. Tv-gan: Generative adversarial network based thermal to visible face recognition. In *2018 international conference on biometrics (ICB)*, pages 174–181. IEEE, 2018.

[15] He Zhang, Vishal M Patel, Benjamin S Riggan, and Shuowen Hu. Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 100–107. IEEE, 2017.

[16] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. Unsupervised generative adversarial cross-modal hashing. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[17] Huijiao Wang, Haijian Zhang, Lei Yu, Li Wang, and Xulei Yang. Facial feature embedded cyclegan for vis-nir translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1903–1907. IEEE, 2020.

[18] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2018.

[19] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3623–3632, 2019.

[20] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13379–13389, 2020.

[21] 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.

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[23] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in neural information processing systems*, pages 406–416, 2017.

[24] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3416, 2018.

[25] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[26] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[27] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[28] Kanglin Liu, Qing Li, and Guoping Qiu. Posegan: A pose-to-image translation framework for camera localization. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:308–315, 2020.

[29] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[30] Y Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Be-yond part models: Person retrieval with refined part pooling. *CoRR*, (1):2–4, 2017.