Image and Vision Computing 94 (2020) 103861

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Coupled generative adversarial network for heterogeneous face recognition

Seyed Mehdi Iranmanesh[a,*], Benjamin Riggan[b], Shuowen Hu[b], Nasser M. Nasrabadi[a]

[a] Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26505, USA
[b] US Army Research Laboratory, Adelphi, MD 20783-1197, USA

ARTICLE INFO

Article history:
Received 12 April 2019
Accepted 3 December 2019
Available online 10 December 2019

Keywords:
Heterogeneous face recognition
Generative adversarial networks
Face verification
Coupled deep neural network
Common latent subspace
Biometrics

ABSTRACT

The large modality gap between faces captured in different spectra makes heterogeneous face recognition (HFR) a challenging problem. In this paper, we present a coupled generative adversarial network (CpGAN) to address the problem of matching non-visible facial imagery against a gallery of visible faces. Our CpGAN architecture consists of two sub-networks one dedicated to the visible spectrum and the other sub-network dedicated to the non-visible spectrum. Each sub-network consists of a generative adversarial network (GAN) architecture. Inspired by a *dense network* which is capable of maximizing the information flow among features at different levels, we utilize a densely connected encoder-decoder structure as the generator in each GAN sub-network. The proposed CpGAN framework uses multiple loss functions to force the features from each sub-network to be as close as possible for the same identities in a common latent subspace. To achieve a realistic photo reconstruction while preserving the discriminative information, we also added a perceptual loss function to the coupling loss function. An ablation study is performed to show the effectiveness of different loss functions in optimizing the proposed method. Moreover, the superiority of the model compared to the state-of-the-art models in HFR is demonstrated using multiple datasets.

1. Introduction

In recent years, there has been significant interest in Heterogeneous Face Recognition (HFR) [1], where the objective is to match visible facial imagery to facial imagery captured in another domain, such as the infrared spectrum [2, 3], polarimetric [4], or millimeter wave [5]. Since there is significantly less facial imagery available in these alternative domains compared to the visible domain, robustness to the variations in wavelength, texture, resolution, noise, and etc., can be difficult to achieve.

The infrared portion of the electromagnetic spectrum can be coarsely divided into reflection-dominated and emission-dominated regions. The reflection-dominated region consists of the following wavelengths: near infrared (NIR; 0.75–1 μm), and shortwave infrared (SWIR; 1–2.5 μm). There has been significant performance improvement in NIR-to-visible face recognition accuracy [6,2] and to some extent, for SWIR-to-visible face recognition accuracy [7, 8]. In Ref. [9], the authors used Restricted Boltzmann Machine (RBM) to learn a common representation for features extracted locally and consequently removed the heterogeneity around each facial point, utilizing Principle Component Analysis (PCA) to obtain the high level features from the local features. In Ref. [10], a novel transductive subspace learning method was proposed for domain invariant feature extraction for VIS-NIR matching problem. Klare et al. [11] used kernel similarities for a set of training subjects as features. Juefei-Xu et al. [12] used a dictionary learning approach to reconstruct images between visible and NIR domains. A common weakness of the prior methods is that they did not use deep non-linear features of face images, which have been shown to produce better results in face recognition problems [13].

The emission-dominated (i.e., thermal) region of infrared spectrum consists of the following wavelengths: midwave infrared (MWIR; 3–5 μm), and longwave infrared (LWIR; 8–12 μm). Thermal facial imagery can be passively acquired without any external illumination because thermal radiation is naturally emitted from facial skin tissue, arising from the underlying vasculature, and other physiological effects. This means MWIR or LWIR imagery is ideal for night-time and low-light scenarios. However, the phenomenological differences between visible and thermal imagery, and the trade-off between wavelength and resolution (or pixel pitch) make matching visible and thermal facial signatures a daunting task.

In recent years, there has been growing research on thermal-to-visible face recognition [11,14-16] and thermal-to-visible detection [17]. Visible images contain rich textural and geometric details

* Corresponding author.
E-mail address: seiranmanesh@mix.wvu.edu (S.M. Iranmanesh).

https://doi.org/10.1016/j.imavis.2019.103861
0262-8856/

across key facial structures (i.e., mouth, eyes, and nose). However, in conventional thermal facial imagery, though some edges around the eyes and eyebrows do appear, but they suffer from significant lack of contrast compared to the corresponding visible images, thus highlighting the large domain gap.

Recently, via an emerging technology [18], the polarization state information of thermal emissions has been exploited to provide additional geometrical and textural details, especially around the nose and the mouth, which complements the textural details of the conventional intensity-based thermal images. This additional information is not available in the conventional intensity-based thermal imaging [18], and is utilized in recent algorithms to enhance thermal-to-visible face recognition [18-20]. Fig. 1 shows a visible face image and its corresponding conventional thermal and polarimetric thermal images.

Algorithms for thermal-to-visible face recognition can be categorized as cross-spectrum feature-based methods, or cross-spectrum image synthesis methods. In cross-spectrum feature-based face recognition a thermal probe is matched against a gallery of visible faces corresponding to the real-world scenario [4], in a feature subspace. The second category synthesizes a visible-like image from a thermal image which can then be used by any commercial visible spectrum face recognition system. Researchers have also investigated a variety of approaches to exploit the polarimetric LWIR thermal images to improve the cross-spectrum face recognition [18,19,21,22]. One of the first methods developed for polarimetric thermal-to-visible face recognition combined the histogram of oriented gradients (HOG) features from S_0, S_1, and, S_2 and combined them together and performed a one-versus-all support vector machine (SVM) classifier to do the face recognition [23]. Another work utilized similar approach to extract features [21]. However, they used partial least square (PLS), on top of the extracted features and learned a one-vs-all PLS discriminant analysis classifier.

Recent cross-spectrum feature based approaches learn a function to explicitly map the polarimetric thermal features to the corresponding visible feature domain representation. Riggan et al. [21] employed deep perpetual mapping (DPM) and coupled neural network (CpNN) [24] for polarimetric thermal-to-visible face recognition. The DPM technique [25] learns a direct mapping between the scale invariant feature transform (SIFT) features from the thermal imagery and the corresponding visible SIFT feature subspace using a multilayer neural network. In contrast, CpNN [24] performs an indirect mapping between thermal and visible SIFT features. The authors in Ref. [24] developed a method to jointly learn two mappings in order to extract the shared latent features. The authors also added one-vs-all PLS classification on top of CpNN or DPM to enhance the recognition accuracy. These two approaches are referred to as PLS$\circ$DPM [4] and PLS$\circ$CpNN [21].

Recently, almost all the state-of-the-art techniques in face recognition have applied deep convolutional neural networks (DCNN) trained on large datasets to construct a compact discriminative feature subspace. This approach also has been applied in other applications such as pedestrian detection [17], and cross-modal retrieval [26] to find a representative embedding subspace. In Ref. [27], the authors trained a network on a private dataset containing 4.4 million labeled images of 4030 different subjects. They also fine-tuned their network with a Siamese network [28] for a face verification task, and extended their work with an expanded dataset which contained 500 million images from 10 million subjects. Sun et al. [29-32] studied a deep neural network architecture employing a joint verification-identification loss function and Bayesian metrics in their works. They used two different datasets, namely, CelebFaces [29] (202,599 images of 10,177 different subjects) and WDRef [33] (99,773 images of 2995 subjects) to train their deep networks. Schroff et al. [34] also trained a deep network using 200 million images of 8 million different subjects. This network gained the best performance on Labeled Faces in the Wild (LFW) [35] dataset, which is a standard unconstrained face recognition benchmark.

The second category of approaches attempt to synthesize a visible-like face image from another modality such as NIR, thermal, or polarimetric thermal input. These methods are beneficial because the synthesized image can be directly utilized by existing face recognition systems developed (i.e., trained) specifically for visible-based facial recognition. Therefore, using this approach one can leverage existing commercial-off-the-shelf (COTS) and government-off-the-shelf (GOTS) solutions. In addition, the synthesized images can be used by human examiners for adjudication purposes. In Ref. [36], the authors developed a method to synthesize a visible-like face image from the polarimetric input. In order to perform synthesis, they utilized DPM to map SIFT features to the corresponding SIFT features in the visible domain, and then reconstructed the visible images from the mapped SIFT features. The authors extended their work in Ref. [37] where they employed a multi-region based approach to jointly optimize the global and local spatial information during the reconstruction. In contrast to the two-step process of Riggan et al. [36,37], Zhang et al. [38] proposed a generative adversarial network (GAN) based approach to reconstruct a more photo-realistic image using multiple loss functions. In additon to GANs optimization, the other non-convex optimization methods have shown great improvements [65]. While cross-spectrum synthesis methods show significant promise, the face recognition performance achieved with synthesis still lags behind the performance of cross-spectrum

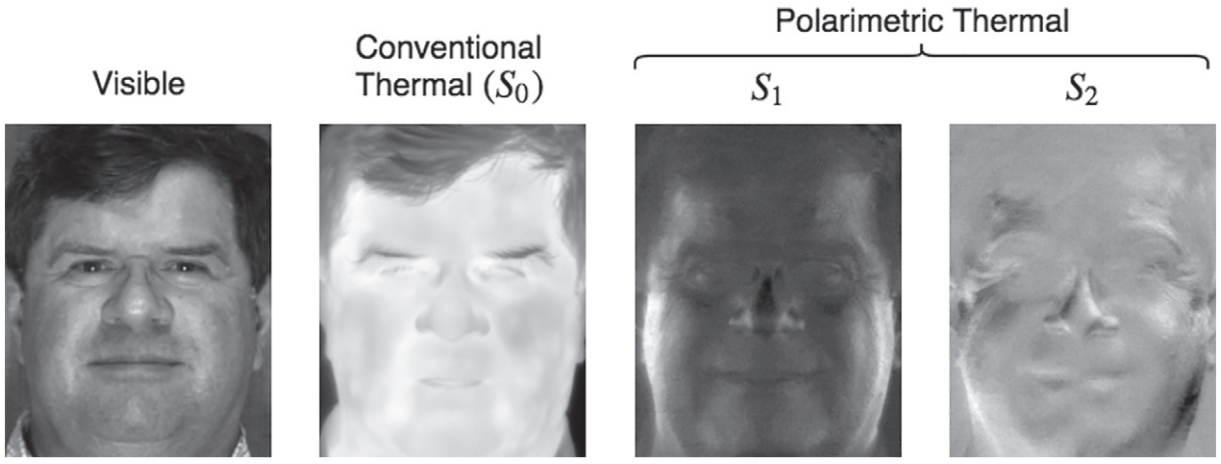

Fig. 1. Visible spectrum and its corresponding conventional thermal (S_0), and polarimetric state information (S_1 and S_2) of a thermal image of a subject.

feature matching based approaches [39]. However, with the constant advancement in GAN architectures and deep generative models, it is expected that synthesis based methods will proceed to outperform the feature-based cross-spectrum matching methods.

Motivated by recent advances in face recognition algorithms using deep approaches and generative models, we propose a novel Coupled Generative Adversarial Network (CpGAN) for cross-spectrum face recognition, which utilizes non-visible modalities to perform a cross-spectrum face recognition task. In Ref. [13], authors used a coupled CNN-based architecture for their face recognition system. However, they evaluated their framework only on near infrared imagery which is visually similar to visible imagery, containing more high frequency details than corresponding thermal imagery. Here, we evaluate the proposed algorithm on different regions of the electromagnetic spectrum from NIR to the more challenging bands such as midwave and longwave infrared. We compare our proposed framework against several different state-of-the-art techniques in the literature such as DPM [25], coupled neural network (CpNN) [24], PLS [16], PLS∘DPM and PLS∘CpNN [4,21]. We present a thorough evaluation using multiple datasets: Wright State (WSRI), Notre Dame X1 (UND X1), Night Vision (NVESD), Polarimetric thermal, and Casia NIR-VIS 2.0 datasets. Our results show that our proposed CpGAN could outperform the existing methods for heterogeneous face recognition.

2. Background

2.1. Polarimetric thermal imagery

In comparison to the conventional thermal imaging that captures intensity-only in the midwave infrared (MWIR) or longwave infrared (LWIR) bands, polarimetric thermal imaging acquires the polarization state information in the thermal infrared spectrum. Polarization states are characterized using the Stokes parameters S_0, S_1, S_2, and S_3, where S0 represents the conventional intensity-only thermal information and S1, S2, and S3 convey polarization state information (see Fig. 1). The polarimetric measurement are made using linear and circular polarizers. The four mentioned Stokes parameters which completely define the polarization states are:

$$S_0 = I^{\circ}_0 + I^{\circ}_{90} \ , \tag{1}$$

$$S_1 = I^{\circ}_0 - I^{\circ}_{90} \ , \tag{2}$$

$$S_2 = I^{\circ}_{45} + I^{\circ}_{-45} \ , \tag{3}$$

$$S_3 = I^{\circ}_R + I^{\circ}_L \ , \tag{4}$$

where $I^{\circ}_0, I^{\circ}_{90}, I^{\circ}_{45}$, and I°_{-45} describe the measured intensity of the light after passing through a linear polarizer with angle of $0^{\circ}, 90^{\circ}, 45^{\circ}$, and -45° related to horizontal axes, respectively. I_R and I_L represent the intensity of the right and left circularly polarized light. Since there is no artificial illumination in passive imaging, there is almost no circularly polarized information in LWIR or MWIR spectrum. Therefore, S_3 is considered to be zero for most of the applications. To quantify the portion of electromagnetic radiation that is linearly polarized, the Degree of Linear Polarization (DoLP), is computed with the linear combination of the Stokes as follows:

$$\text{DoLP} = \frac{\sqrt{S_1^2 + S_2^2}}{S_0} \ . \tag{5}$$

2.2. DenseNets

Traditional convolutional feed-forward networks such as VGG [40], connect the output of the l^{th} layer as the input to the next layer, which is equal to the following transition: $x_l = H_l(x_{l-1})$, where H_l is the convolutional mapping from $l-1$ to l. In Resnet [41], authors made a change in this transition information by adding a skip-connection which bypasses the non-linear transformation with an identity function:

$$x_l = H_l(x_{l-1}) + x_{l-1} \ . \tag{6}$$

A benefit of Resnet architecture is that through the identity function, the gradient of the cost function can progress directly from later layers to the earlier layers. However, the combination of the identity function and output of H_l might prevent the information flow in the network [42].

In order to improve the information flow between different layers, in Densenet [42] authors provided a different connectivity between different layers in which there is a direct connection between any layer and all the subsequent layers. Therefore, the l^{th} layer receives the feature maps of all the previous layers, $x_0, x_1, \ldots, x_{l_{-1}}$ as input:

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]) \ , \tag{7}$$

where $[x_0, x_1, \ldots, x_{l_{-1}}]$ represents the concatenation of the feature maps produced from the previous layers $0, \ldots, l-1$ [42] (see dense block in Fig. 2).

2.3. Generative adversarial networks

The generative adversarial network consists of two sub-networks, namely a generator and a discriminator which compete with each other in a minimax game. For the generator to learn the distribution p_g over the data x, the authors consider a prior on the input noise variables $p_z(z)$ [43]. Generator network G is a differentiable function with a parameter θ_g which performs a mapping to the data space $G(z; \theta_g)$. On the other hand, the discriminator network is also a differentiable function $D(.; \theta_d)$ which performs a binary classification between the real data x and the generated data $G(z)$. At the same time, network G tries to fool the discriminator by minimizing $\log(1 - D(G(z)))$. In other words, D and G play a two-player minimax game which resembles minimizing the Jenson-Shannon divergence [43] as follows:

$$\min_G \max_D E_{x \sim P_{data(x)}}[\log D(x)] + E_{z \sim P_z}[\log(1 - D(G(z)))]. \tag{8}$$

2.4. Conditional generative adversarial networks

Conditional adversarial networks is an extension of generative adversarial networks in which both the generator and discriminator are conditioned on some auxiliary information y. The extra information y can be any kind of information such as class label or other modalities data. The objective of the conditional GAN is the same as the classical GAN. The only exception is that in the conditional GAN both the discriminator and generator are conditioned on the auxiliary information as follows [44]:

$$\min_G \max_D E_{x \sim P_{data(x)}}[\log D(x|y)] + E_{z \sim P_z}[\log(1 - D(G(z|y)))], \tag{9}$$

3. Proposed method

The proposed CpGAN is illustrated in Fig. 3. The proposed approach consists of two generators and two discriminators which are coupled with each other. In the following subsections, we explain these modules in detail.

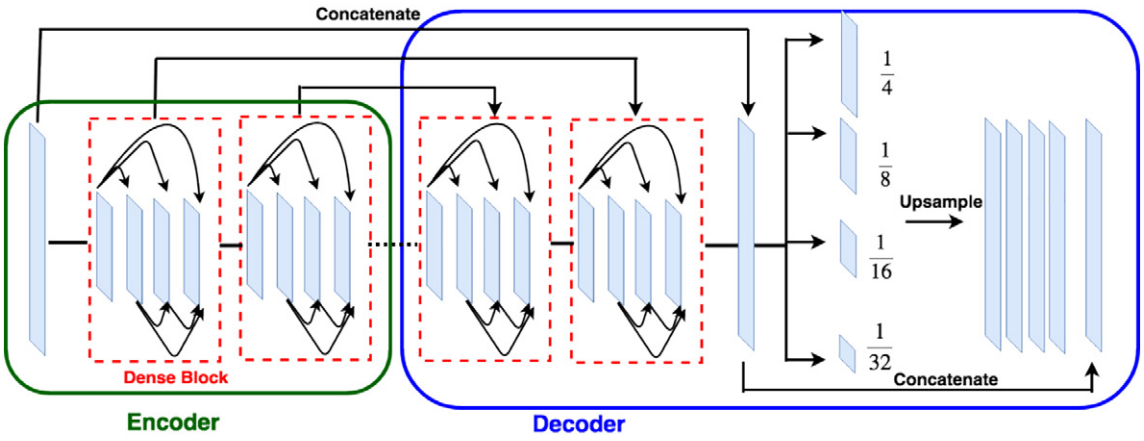

Fig. 2. An overview of the pyramid densely connected network.

3.1. Pyramid densely connected network

This network is a densely connected encoder-decoder structure which utilizes the features from multi layers of a CNN [45]. In this framework, a dense block [42] is used as the basic structure since it can maximize the information flow and has better convergence by connecting all the layers. The encoder part of the network consists of three dense blocks with their corresponding down-sampling operations which shrinks the feature map to 1/32 of the input size. The decoder part is responsible for reconstructing the original size image from the embedding subspace and it stacks five dense blocks with the refined up-sampling transition blocks [46,47]. Moreover, the concatenations are performed on the feature maps with the same size. Inspired by the use of global context information in classification and segmentation, this network tries to capture more global information, using multi-level pyramid pooling blocks [48,49]. This operation is done to make sure that features from different scales are embedded in the final result. Therefore, four different operations with pooling sizes of 1/32, 1/16, 1/8, and 1/4 is selected. All the four level features are up-sampled to the original size and are concatenated together. Fig. 2 illustrates the overview of the pyramid densely connected network.

3.2. Deep cross-modal face recognition

The overall objective of the proposed model is identification of non-visible faces which were not seen in the training phase. For this reason, we couple two pyramid densely connected networks one dedicated to the visible spectrum (Vis-GAN) and the other one to the non-visible spectrum (NVis-GAN). Each network performs a non-linear transformation of the input space. The final objective of our proposed CpGAN is to learn a joint, deep embedding that captures the interrelationship between the visible and non-visible facial imagery for spectrally invariant face recognition. In order to find a common latent embedding subspace between these two different domains, we couple two pyramid densely connected networks (Vis-GAN and NVis-GAN) via a contrastive loss function [28].

The contrastive loss function (ℓ_{cont}) encourages the genuine pairs (i.e., visible and non-visible images with faces of corresponding subjects) to be "close" in terms of some metric (usually the euclidean

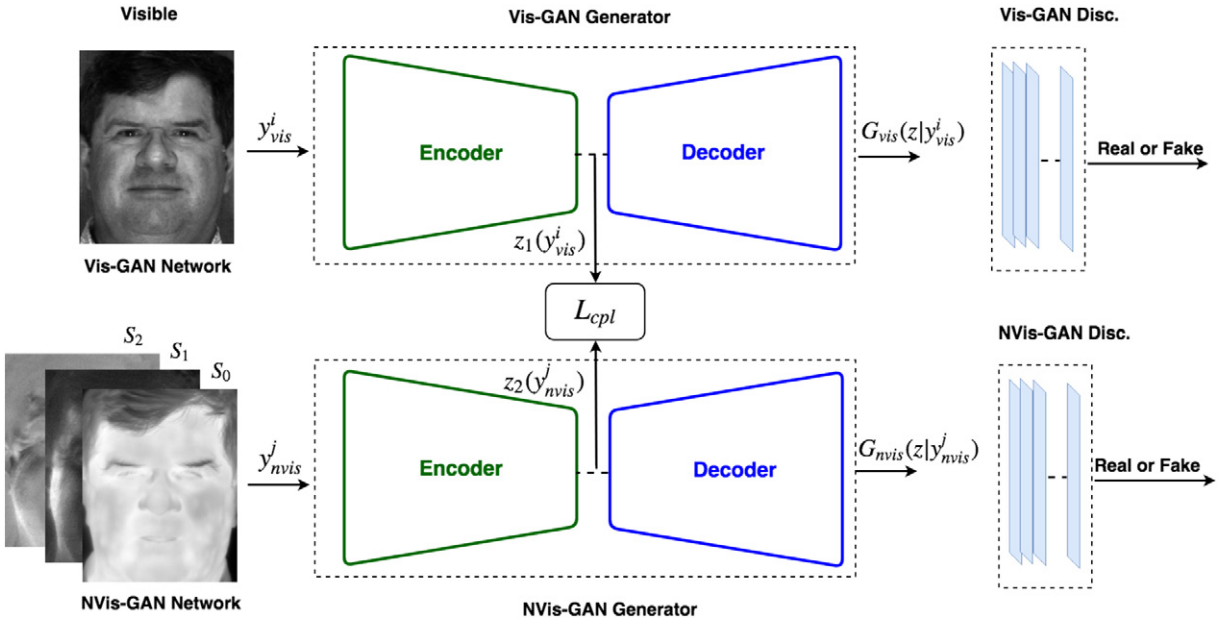

Fig. 3. Proposed network using two GAN based sub-networks (Vis-GAN and NVis-GAN) coupled by contrastive loss function. Here, the input to NVIS-GAN is polarimetric data (S_0, S_1, S_2). In the case of other non-visible modalities such as (NIR, MWIR, and LWIR), the framework remains the same and only the input to the NVIS-GAN is changed accordingly.

distance) and the impostor pairs (i.e., visible and non-visible images containing faces of different subjects) to be distant from each other (see VisGAN and NVis-GAN networks at their bottlenecks in Fig. 3). Similar to Ref. [28], our contrastive loss is of the form:

$$\ell_{cont}(z_1(y^i_{vis}), z_2(y^j_{nvis}), y_{cont}) = (1 - y_{cont})L_{gen}(d(z_1(y^i_{vis}), z_2(y^j_{nvis})) + y_{cont}L_{imp}(d(z_1(y^i_{vis}), z_2(y^j_{nvis})), \quad (10)$$

where y^i_{vis} is the input for the Vis-GAN (i.e., visible face image), and y^j_{nvis} is the input for the NVis-GAN (i.e., non-visible face images). y_{cont} is a binary label, L_{gen} and L_{imp} represent the partial loss functions for the genuine and impostor pairs, respectively, and $d(z_1(y^i_{vis}), z_2(y^j_{nvis}))$ indicates the Euclidean distance between the embedded data in the embedded common feature subspace. $z_1(.)$ and $z_2(.)$ are the deep convolutional neural network based embedding functions, which transform y^i_{vis} and y^j_{nvis} into a common latent embedding subspace, respectively. The binary label, y_{cont}, is assigned a value of 0 when both modalities, i.e., visible and non-visible, form a genuine pair, or, equivalently, the inputs are from the same class ($cl^i = cl^j$). On the contrary, when the inputs are from different classes, which means they form an impostor pair, y_{cont} is equal to 1. In addition, L_{gen} and L_{imp} are defined as follows:

$$L_{gen}(d(z_1(y^i_{vis}), z_2(y^j_{nvis}))) = \frac{1}{2} \times ||z_1(y^i_{vis}) - z_2(y^j_{nvis})||_2^2 \quad \text{for} \quad cl^i = cl^j, \quad (11)$$

and

$$L_{imp}(d(z_1(y^i_{vis}), z_2(y^j_{nvis}))) = \frac{1}{2} \times \max(0, m - ||z_1(y^i_{vis}) - z_2(y^j_{nvis})||_2^2) \quad \text{for } cl^i \neq cl^j, \quad (12)$$

where m is the contrastive margin. The coupling loss function can be written as:

$$L_{cpl} = 1/N^2 \sum_{i=1}^{N}\sum_{j=1}^{N} \ell_{cont}(z_1(y^i_{vis}), z_2(y^j_{nvis}), y_{cont}), \quad (13)$$

where N is the number of samples. It should be noted that the contrastive loss function Eq. (13) considers the subjects' labels implicitly. Therefore, it has the ability to find a discriminative embedding space by employing the data labels in contrast to some other metrics such as the Euclidean distance. This discriminative embedding space is useful in identifying a non-visible probe photo against a gallery of visible photos.

3.3. Generative adversarial loss

Let G_{vis} and G_{nvis} denote the generators that synthesize a visible image from an input visible and a non-visible image, respectively. To synthesize the output and to make sure that the synthesized images generated by the two generators are indistinguishable from the corresponding ground truth visible image, we utilized the GAN loss function in Ref. [44]. As it is shown in Fig. 3, the first generator G_{vis} is responsible to generate a visible image when the network is conditioned on a visible image. On the other hand, the second generator G_{nvis} tries to generate the same visible image from the non-visible image which has a more challenging task compared to the first generator. Therefore, the total loss for the coupled GAN is as follows:

$$L_{GAN} = L_{vis} + L_{nvis}, \quad (14)$$

where the GAN loss function related to the Vis-GAN is given as:

$$L_{vis} = \min_{G_{vis}} \max_{D_{vis}} E_{x^i \sim P_{vis(x)}}[\log D(x^i|y^i_{vis})] + E_{z\sim P_z}[\log(1 - D(G(z|y^i_{vis})))], \quad (15)$$

where y^i_{vis} is the visible image used as condition for the Vis-GAN and x^i is the real data. It should be noted that for the Vis-GAN the real data x^i and the condition y^i_{vis} are both visible. Similarly, the loss for the NVis-GAN is given as:

$$L_{nvis} = \min_{G_{nvis}} \max_{D_{nvis}} E_{x^j \sim P_{vis(x)}}[\log D(x^j|y^j_{nvis})] + E_{z\sim P_z}[\log(1 - D(G(z|y^j_{nvis})))], \quad (16)$$

where y^j_{nvis} is the non-visible image used as condition for the NVis-GAN and x^j is the real data (which is visible). It should be noted that x^i is the same as x^j if they refer to the same subject ($cl^i = cl^j$), otherwise they are not the same.

3.4. Overall loss function

The proposed approach contains the following loss function: the Euclidean $L_{E_{vis}}$ and $L_{E_{nvis}}$ losses which are enforced on the recovered visible images from the Vis-GAN and NVis-GAN networks, respectively, are defined as follows:

$$L_{E_{vis}} = ||G_{vis}(z|y^i_{vis}) - x^i||_2^2, \quad (17)$$

$$L_{E_{nvis}} = ||G_{nvis}(z|y^j_{nvis}) - x^j||_2^2, \quad (18)$$

$$L_E = L_{E_{vis}} + L_{E_{nvis}}. \quad (19)$$

The L_{GAN} (Eq. (14)) loss is also added to generate sharper images. In addition, based on the success of perceptual loss in low-level vision tasks [50,51], the perceptual loss is added to the NVis-GAN to preserve more photo realistic details as follows:

$$L_{P_{nvis}} = \frac{1}{C_p W_p H_p} \sum_{c=1}^{C_p}\sum_{w=1}^{W_p}\sum_{h=1}^{H_p} ||V(G_{nvis}(z|y^j_{nvis}))^{c,w,h} - V(x^j)^{c,w,h}||, \quad (20)$$

where x^j is the ground truth visible image, $G_{nvis}(z|y^j_{nvis})$ is the output of NVis-GAN generator. V(.) represents a non-linear CNN transformation and C_p, W_p, H_p are the dimension of a particular layer in V. It should be noted that the perceptual loss is just used in the NVis-GAN.

Finally, the contrastive loss function Eq. (13) is added to train both networks Vis-GAN and NVis-GAN jointly to make the embedding space of the mentioned networks as close as possible and to preserve a more discriminative and distinguishable shared space. Therefore, the total loss function for the proposed CpGAN is as follows:

$$L_T = L_{cpl} + \lambda_1 L_E + \lambda_2 L_{GAN} + \lambda_3 L_{Pnvis}, \quad (21)$$

where L_{cpl} is the coupling loss (Eq. (13)) term which is the contrastive loss function, the second is the total $L2$ loss for the Vis-GAN and NVis-GAN. L_{GAN} and $L_{P_{nvis}}$ are the GAN, and perceptual loss functions for the Vis-GAN, respectively. λ_1, λ_2, and λ_3 are the hyper-parameters which weight the Euclidean, the adversarial, and the perceptual losses, respectively.

3.5. Testing phase

During the testing phase, only the NVis-GAN is used. For a given test probe y^t_{nvis}, NVis-GAN is employed in the proposed CpGAN to

synthesize the visible image $G_{nvis}(z|y^t_{nvis}) = \hat{x}^t_{vis}$. Eventually, the identification of face recognition is done, by calculating the minimum Euclidean distance between the synthesized image and visible gallery images as follows:

$$x^{t^*}_{vis} = \underset{x^t_{vis}}{\text{argmin}} \quad ||x^t_{vis}, \hat{x}^t_{vis}|| \,, \tag{22}$$

where $\hat{x}^t_{vis}$ is the synthesized probe face image and $x^{t^*}_{vis}$ is the selected matching visible face image within the gallery of face images.

4. Experiments and results

4.1. Implementation details

The network is trained on a Nvidia Titan X GPU using the PyTorch framework. We choose $\lambda_3 = 0.5$ and $\lambda_{1,2} = 1$. For training, we used the Adam optimizer [52] with a first-order momentum of 0.5 and a learning rate of 0.0002 and a batch size of 4. The perceptual loss is assessed on relu3-1 layer of a pre-trained VGG [40] model for the Imagenet dataset [53].

4.2. Heterogeneous face recognition datasets

In order to evaluate the proposed CpGAN model, we utilize six different heterogeneous face recognition databases:

1) Wright State (WSRI) [54],
2) Notre Dame X1 (UND X1) [55],
3) Night Vision (NVESD) [56],
4) Casia NIR-VIS 2.0 [57],
5) Casia HFB [58],
6) Polarimetric thermal [4],

in order to test the NIR-to-visible, MWIR-to-visible, LWIR-to-visible and polarimetric thermal-to-visible face recognition applications. Table 1 provides an overview of the datasets used in this work — each database is briefly described below:

WSRI dataset consists of 1615 visible and 1615 MWIR images from 64 different identities. There are approximately 25 images per subject approximately with different facial expressions. The original resolution of the visible images is 1004 × 1004, and 640 × 512 for the MWIR modality. After preprocessing, the images from both modalities are resized to 235 × 295 pixels. This database is split randomly into a set of 10 subjects for training set and remaining 54 subjects for testing set.

UND X1 dataset contains LWIR and visible images related to 241 subjects with different variations in lighting, expression and time lapse. The original resolutions of the images are 1600 × 1200 pixels for the visible modality and 320 × 240 pixels for the LWIR modality. Both modalities are resampled to 150 × 110 pixels after preprocessing.

The training set composed of 159 subjects captured in the visible and LWIR modalities with only one image per subject. On the other hand, the test set contains the remaining 82 subjects with multiple images per subject. This database is challenging due to the low resolution and noise present in the LWIR imagery. This leads to significant difference between the two modalities in this dataset.

NVESD dataset is collected by the U.S. Army CERDEC-NVESD in 2012 from 50 different subjects. The dataset composed of 450 images in each modality. The images were captured simultaneously from different identities with the original resolution of 640 × 480 pixels for all of the modalities. After preprocessing as in Ref. [16], the image resolution is resampled to 174 × 174 and dataset is split into training and testing sets.

CASIA NIR-VIS 2.0 dataset contains the visible and NIR images from 725 different identities. The images were not captured simultaneously. For each subject, there are 1–22 visible images and 5–50 NIR images with different expressions, poses, glasses, and distance to camera/sensor. The original resolution of the images for both modalities are 640 × 480 pixels. After preprocessing, the cropped image sizes are 128 × 128. This database provides a part of data for the sake of parameter tuning, and 10 remaining parts for reporting the experimental results.

CASIA HFB dataset contains 202 subjects. Similar to the CASIA NIR-VIS 2.0, this dataset has two views where the first view is for parameter selection and View2 is for the sake of evaluation. This dataset contains about 1000 visible images and 1500 NIR images for training and similarly 1000 visible and 1500 NIR images for testing. The resolution of the images before and after preprocessing is the same as the NIR-VIS 2.0 dataset.

Polarimetric Thermal Face dataset [4] contains polarimetric LWIR and visible face images of 60 subjects. Data was collected at three different distances: Range 1 (2.5 m), Range 2 (5 m), and Range 3 (7.5 m). At each range, baseline and expressions data were collected. In the baseline condition, the subject was asked to keep a neutral expression looking at the polarimetric thermal sensor. On the other hand, in the expression condition, the subject was asked to count numerically upwards from one, resulting in different expressions in the mouth to eye regions. Each subject has 16 images of visible and 16 polarimetric LWIR images in which four images are from the baseline condition and the remaining 12 images are from the expression condition.

Table 1
Summary of heterogeneous face recognition datasets used for comparing models.

Database	Source	Target	# subjects	Variations
WSRI	Visible	MWIR	64	E
UND X1	Visible	LWIR	241	E
NVESD	Visible	MWIR & LWIR	50	E,D
Casia NIR-VIS 2.0	Visible	NIR	725	P,E,G,D
Casia HFB	Visible	NIR	202	P,E,G,D
Polarimetric thermal	Visible	S_0, S_1, S_2	60	E,D

4.3. WSRI and UND results

The network for the visible face images (Vis-GAN) and the network for the non-visible face images (NVis-GAN) have the same structure. These images are resized to 256 × 256 before passing to the network. To benefit from the pre-defined weights of the DenseNet [42], the first convolutional layer and the first three Dense-blocks have been leveraged from a pre-trained DenseNet 121 as the encoder structure. At the end of the encoder part where the feature map size is 1/32 of the original input spatial dimensions, the two sub-networks (Vis-GAN and NVis-GAN) are coupled together via a contrastive loss function (see Fig. 3) to construct the CpGAN framework.

To increase the correlation between the two modalities of visible and LWIR (UND X1 and NVESD datasets), each modality was preprocessed. We applied difference of Gaussians (DoG) filter, to emphasize the edges in addition to removing high and low frequency noise. The DoG filter which is the difference of two Gaussian kernels with different standard deviations is defined as follows:

$$DG(I, \sigma_0, \sigma_1) = [g(x, y, \sigma_0) - g(x, y, \sigma_1)] * I(x, y) \,, \tag{23}$$

where DG is the DoG filtered image, $*$ is the convolution operator, and g is the Gaussian kernel which is defined in:

$$g(x, y, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}} \,. \tag{24}$$

The training set is used to transform the visible and non-visible features to a shared latent embedding subspace. Also at the same time, the network tries to synthesize visible modality from the shared latent subspace in the GAN framework. To train the network, the genuine and impostor pairs are constructed. The genuine pair is constructed from the same subject images in two different modalities. For the impostor pair, a different subject is selected for each modality. In general, the number of the generated impostor pairs are significantly larger than the genuine pairs. For the sake of balancing the training set, we consider the same number of genuine and impostor pairs. After training the network, during the testing phase, only the NVis-GAN sub-network is used for the evaluation. For a given probe, the network is used to synthesize the visible image. Afterwards, the Euclidean distance is used to match the synthesize image to its closest image from the gallery. The ratio of the number of correctly classified subjects and the entire number of subjects is computed as the identification rate.

The identification rate of our proposed approach for both WSRI and UND X1 datasets is reported in Table 2. In addition, we compare the performance of our method with some state-of-the-art methods in the literature such as CpNN [24], PLS [14], bilevel coupled dictionary learning (BCDL) [59], and kernel bilevel coupled dictionary learning (K-BCDL) [24]. The tabulated results show the improved performance of the proposed method and its effectiveness in synthesizing the visible modality from the non-visible modality.

4.4. NVESD results

We compare our proposed CpGAN with the reported results in the literature on the NVESD dataset. For the sake of comparison, we perform the same split as in Ref. [24] on the dataset for the train and test set. Therefore, we train our proposed framework on training set with 10 subjects and report the rank-1 classification performance on the test set of 40 subjects. This database contains two different non-visible modalities, namely, MWIR and LWIR. Table 3 shows the reported results of our proposed method and as well as the other state-of-the-art models. As it is shown in Table 3, our proposed method performance surpasses the other methods in the literature for both MWIR-to-visible and LWIR-to-visible face recognition.

4.5. CASIA results

In this experiment, we compare our results with the results reported in Ref. [60]. For the sake of fair comparison, we perform the same set of experiments as in Ref. [60]. The dataset has two views in which View1 is used for parameter tuning and View2 with 10 different splits are used for testing. Number of images in HFB dataset is about 1000 visible images and 1500 NIR images during the testing phase. The CASIA NIR-VIS 2.0 restricts algorithms to one gallery per subject during the testing phase. Therefore, there are only 358 gallery images for the comparison, while there are about 6000 probe NIR images for testing.

In addition to the higher number of images in NIR-VIS 2.0, some of the images in this dataset contain more challenging non-frontal poses, while the HFB images were taken in a more controlled environment. Moreover, the restriction of one image per gallery subject, makes the NIR-VIS 2.0 dataset more challenging. Tables 4 and 5 show the results of the proposed method compared to the other methods in the literature for the NIR-VIS 2.0 and HFB datasets, respectively. Following Ref. [60], the reported result is the average of 10 different experimental setups. The results show that our method performs

Table 2
Rank-1 identification rates of the proposed method and the baseline methods for WSRI and UND X1 datasets.

Method	WSRI	UND X1
PLS	83.7%	41.0%
BCDL	93.1%	50.5%
K-BCDL	95.9%	52.0%
CpNN	97.2%	51.9%
CpGAN	**97.8%**	**76.4%**

Table 3
Rank-1 identification rates of the proposed method and the baseline methods for MWIR and LWIR on NVESD dataset.

Method	MWIR	LWIR
PLS	82.4%	70.4%
BCDL	90.7%	90.6%
K-BCDL	93.3%	92.5%
CpNN	94.4%	89.1%
CpGAN	**96.1%**	**93.9%**

Table 4
Performance comparison to other baselines on View2 of CASIA NIR-VIS 2.0 dataset.

NIR-VIS 2.0	Rank 1	Std. Dev.	FAR = 0.001
CpNN	33.1%	6.6	76.35
C-CBFD [61]	81.8%	2.3	47.3
[62]	85.9%	0.9	78.0
[9]	86.2%	0.98	81.3
[63]	95.74%	0.52	91.03
[60]	92.6%	0.64	81.6
CpGAN	**96.63%**	0.56	87.05

Table 5
Performance comparison to other baselines on View2 of CASIA HFB 2.0 dataset.

HFB	Rank 1	FAR = 0.01	FAR = 0.001
CpNN	39.8%	84.4	72.49
IDNet [13]	80.9%	70.4	36.2
P-RS [11]	87.8%	98.2	95.8
C-DFD [64]	92.2%	85.6	65.5
THFM [10]	99.28%	99.66	98.42
[9]	99.38%	–	92.25
[60]	99.52%	98.6	91.8
CpGAN	**99.64%**	98.4	89.7

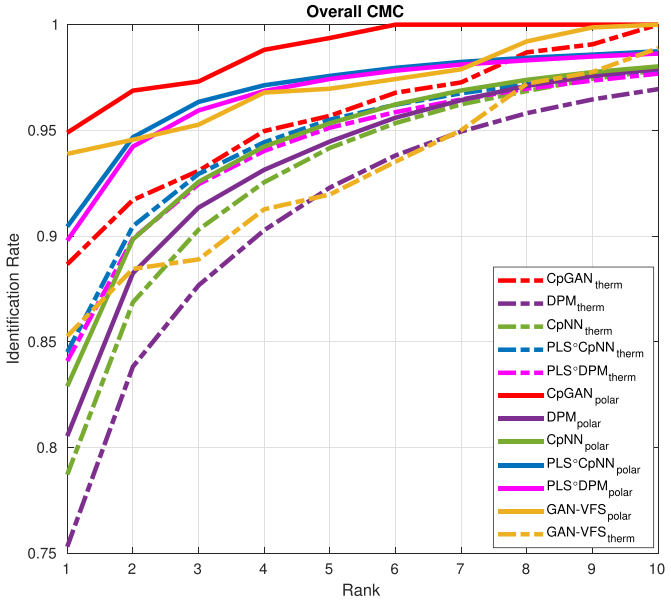

Fig. 4. Overall CMC curves from testing PLS, DPM, CpNN, PLS∘DPM, PLS∘CpNN, GAN-VFS, and CpGAN using polarimetric and thermal probe samples, matching against a visible spectrum gallery.

Table 6
Rank-1 identification rate for cross-spectrum face recognition using polarimetric thermal and conventional thermal (S_0) probe imagery.

Scenario	Rank-1 Identification Rate							
	Probe	PLS	DPM	CpNN	PLS∘DPM	PLS∘CpNN	GAN-VFS	CpGAN
Overall	Polar	0.5867	0.8054	0.8290	0.8979	0.9045	0.9382	**0.9549**
	Therm	0.5305	0.7531	0.7872	0.8409	0.8452	0.8561	**0.8905**
Expressions	Polar	0.5658	0.8324	0.8597	0.9565	0.9559	0.9473	**0.9684**
	Therm	0.6276	0.7887	0.8213	0.8898	0.8907	0.8934	**0.9176**
Range 1 Baseline	Polar	0.7410	0.9092	0.9207	0.9646	0.9646	0.9653	**0.9867**
	Therm	0.6211	0.8778	0.9102	0.9417	0.9388	0.9412	**0.9637**
Range 2 Baseline	Polar	0.5570	0.8229	0.8489	0.9105	0.9187	0.9263	**0.9659**
	Therm	0.5197	0.7532	0.7904	0.8578	0.8586	0.8701	**0.8993**
Range 3 Baseline	Polar	0.3396	0.6033	0.6253	0.6445	0.6739	0.8491	**0.8987**
	Therm	0.3448	0.5219	0.5588	0.5768	0.6014	0.7559	**0.7912**

very well compared to the other methods on the NIR-VIR 2.0 dataset which is more challenging. Moreover, since many other methods have been developed for NIR and evaluated on the HFB dataset, the improvement of 1% in Rank-1 identification performance achieved by the proposed algorithm is significant.

4.6. Polarimetric thermal results

For the polarimetric thermal face dataset, we consider the same CpGAN architecture. We pass S_0, S_1, and S_2 to the NVis-GAN's three channels as the input as shown in Fig. 3.

In each experiment, the dataset is partitioned randomly into the training and testing sets. The same set of training and testing data is used to evaluate PLS, DPM, CpNN, PLS∘DPM, PLS∘CpNN, GAN-VFS [38], and the proposed CpGAN network. Fig. 4 shows the overall cumulative matching characteristics (CMC) curves for our proposed method and the other state-of-the-art methods over all the three different data ranges as well as the expressions data at Range 1. For the sake of comparison, in addition to the polarimetric thermal-to-visible face recognition performance, Fig. 4 also shows the results for the conventional thermal-to-visible face recognition for some of the methods, namely PLS, PLS∘DPM, PLS∘CpNN, CpNN, and CpGAN. For conventional thermal-to-visible face recognition, all the mentioned methods follow the same procedure as before, except only using the S_0 Stokes image. Fig. 4 illustrates that exploiting the polarization information of the thermal spectrum enhances the cross-spectrum face recognition performance compared to using the conventional intensity-only information alone. Fig. 4 also shows the superior performance of our approach compared to the state-of-the-art methods. In addition, our method could achieve perfect accuracy at Rank-5 and above.

Table 6 tabulates the Rank-1 identification rates for five different scenarios: overall (which corresponds to Fig. 4), Range 1 expressions, Range 1 baseline, Range 2 baseline, and Range 3 baseline. In our proposed approach, exploiting polarization information enhance the Rank-1 identification rate by 1.87%, 5.13%, 4.49%, and 5.92% for Range 1 baseline, Range 1 expression, Range 2 baseline, and Range 3 baseline compared to conventional thermal-to-visible face recognition. This table reveals that using deep coupled generative adversarial network technique with the contrastive loss function to transform different modalities into a distinctive common embedding subspace is superior to the other embedding techniques such as PLS∘CpNN. It also shows the effectiveness of our method in exploiting polarization information to improve cross-spectrum face recognition performance.

5. Ablation study

In order to illustrate the effect of adding different loss functions and their improvement in our proposed framework, we perform a study with the following evaluations using the polarimetic dataset: 1) Polar-to-visible using the coupled framework with using only $L_{cpl} + L_E$ loss, 2) Polar-to-visible using the proposed framework with $L_{cpl} + L_E + L_{GAN}$ loss functions, and 3) Polar-to-visible with all the loss functions in the proposed framework (Eq. (21)). Fig. 5 shows the reconstruction results for a random subject in this dataset. We can conclude from Fig. 5 (c), that using $L_{cpl} + L_E$ loss results in a blurry image with reduced high frequency details. However, adding L_{GAN} loss function (Eq. (14)) to the framework leads to a sharper and more vivid images. Moreover, by adding the perceptual loss to the NVis-GAN sub-network, the results become more visually pleasing by removing some artifacts added by L_{GAN}.

For better understanding of different loss functions and their effect on the proposed framework results, we plot the receiver operation characteristic (ROC) curves corresponding to the mentioned three different settings of the framework. As it is shown in Fig. 6 the L_{GAN} has an important rule in the enhancement of our proposed approach. Also, adding a perceptual loss enhances the face recognition performance as well as generating visually more realistic images.

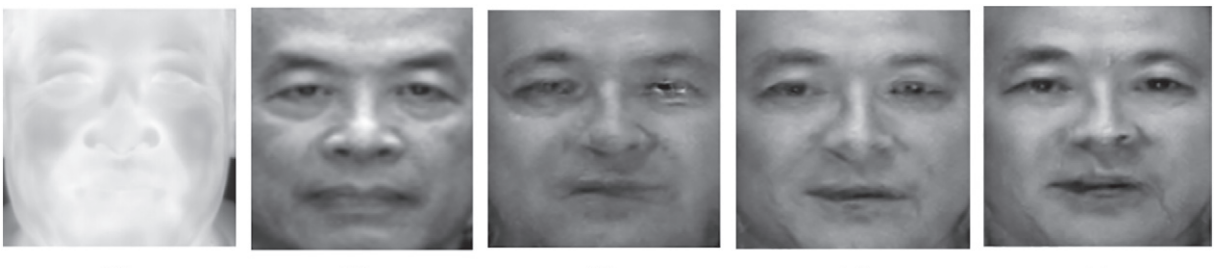

Fig. 5. Comparison of visible face images synthesized with different experimental configurations. (a) Raw polarimetric image (S_0 is just shown in here). (b) Ground truth visible images. (c) Reconstructed images with $L_{cpl} + L_E$. (d) Reconstructed images with $L_{cpl} + L_E + L_{GAN}$. (e) Reconstructed images with CpGAN (Eq. (21)).

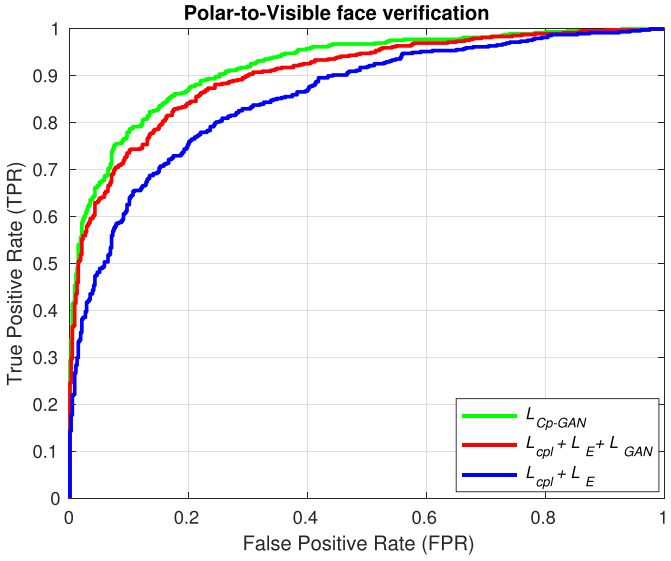

Fig. 6. The ROC curves corresponding to the ablation study.

6. Conclusion

In this work, we proposed a coupled generative adversarial network to synthesize visible image from a non-visible image for the heterogeneous face recognition task. The CpGAN contains two GAN based sub-networks dedicated to visible and non-visible input images. The proposed network is capable of transforming the visible and non-visible modalities into a common discriminative embedding subspace and subsequently synthesizing the visible images from that subspace. In order to efficiently synthesize a realistic visible image from the non-visible modality, a densely connected encoder-decoder structure is used as the generator in each sub-network. An ablation study was performed to demonstrate the enhancement obtained by different losses in the proposed method. The experiments on different HFR datasets with different range of electromagnetic spectrum showed the effectiveness of the proposed method compared to the other state-of-the-art methods. The results also revealed that the proposed framework could exploit polarimetric thermal information to enhance the thermal-to-visible face recognition performance.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Ouyang, T. Hospedales, Y.-Z. Song, X. Li, A Survey on Heterogeneous Face Recognition: Sketch, Infra-red, 3D and Low-resolution,arXiv preprint arXiv:1409.5114. 2014.
[2] B. Klare, A.K. Jain, Heterogeneous face recognition: matching NIR to visible light images, 20th International Conference on Pattern Recognition (ICPR), 2010, IEEE. 2010, pp. 1513–1516.
[3] F. Nicolo, N.A. Schmid, Long range cross-spectral face recognition: matching SWIR against visible light images, IEEE Trans. Inf. Forensics Secur. 7 (6) (2012) 1717–1726.
[4] S. Hu, N.J. Short, B.S. Riggan, C. Gordon, K.P. Gurton, M. Thielke, P. Gurram, A.L. Chan, A polarimetric thermal database for face recognition research, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016. pp. 119–126.
[5] E. Gonzalez-Sosa, R. Vera-Rodriguez, J. Fierrez, V.M. Patel, Millimetre wave person recognition: hand-crafted vs learned features, IEEE International Conference on Identity, Security and Behavior Analysis (ISBA), 2017, IEEE. 2017, pp. 1–7.
[6] D. Yi, R. Liu, R. Chu, Z. Lei, S.Z. Li, Face matching between near infrared and visible light images, International Conference on Biometrics, Springer. 2007, pp. 523–530.
[7] T. Bourlai, N. Kalka, A. Ross, B. Cukic, L. Hornak, Cross-spectral face verification in the short wave infrared (SWIR) band, 20th International Conference on Pattern Recognition (ICPR), 2010, IEEE. 2010, pp. 1343–1347.
[8] F. Nicolo, N.A. Schmid, Long range cross-spectral face recognition: matching SWIR against visible light images, IEEE Trans. Inf. Forensics Secur. 7 (6) (2012) 1717–1726.
[9] D. Yi, Z. Lei, S.Z. Li, Shared representation learning for heterogenous face recognition, 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015, vol. 1, IEEE. 2015, pp. 1–7.
[10] J.-Y. Zhu, W.-S. Zheng, J.-H. Lai, S.Z. Li, Matching NIR face to VIS face using transduction, IEEE Trans. Inf. Forensics Secur. 9 (3) (2014) 501–514.
[11] B.F. Klare, A.K. Jain, Heterogeneous face recognition using kernel prototype similarities, IEEE Trans. Pattern Anal. Mach. Intell. 35 (6) (2013) 1410–1422.
[12] F. Juefei-Xu, D.K. Pal, M. Savvides, NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction, Proceedings of the IEEE conference on Computer Vision and Pattern Recognition workshops, 2015. pp. 141–150.
[13] C. Reale, N.M. Nasrabadi, H. Kwon, R. Chellappa, Seeing the Forest from the Trees: A Holistic Approach to Near-infrared Heterogeneous Face Recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016. pp. 54–62.
[14] J. Choi, S. Hu, S.S. Young, L.S. Davis, Thermal to Visible Face Recognition, Maryland Univ College Park. 2012.
[15] T. Bourlai, A. Ross, C. Chen, L. Hornak, A study on using mid-wave infrared images for face recognition, Proc. SPIE, 8371, 2012. pp. 83711K.
[16] S. Hu, J. Choi, A.L. Chan, W.R. Schwartz, Thermal-to-visible face recognition using partial least squares, JOSA A 32 (3) (2015) 431–442.
[17] D. Xu, W. Ouyang, E. Ricci, X. Wang, N. Sebe, Learning Cross-modal Deep Representations for Robust Pedestrian Detection,arXiv preprint arXiv:1704.02431. 2017.
[18] K.P. Gurton, A.J. Yuffa, G.W. Videen, Enhanced facial recognition for thermal imagery using polarimetric imaging, Opt. Lett. 39 (13) (2014) 3857–3859.
[19] N. Short, S. Hu, P. Gurram, K. Gurton, A. Chan, Improving cross-modal face recognition using polarimetric imaging, Opt. Lett. 40 (6) (2015) 882–885.
[20] N. Short, S. Hu, P. Gurram, K. Gurton, Exploiting polarization-state information for cross-spectrum face recognition, IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2015, IEEE. 2015, pp. 1–6.
[21] B.S. Riggan, N.J. Short, S. Hu, Optimal feature learning and discriminative framework for polarimetric thermal to visible face recognition, IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, IEEE. 2016, pp. 1–7.
[22] A.J. Yuffa, K.P. Gurton, G. Videen, Three-dimensional facial recognition using passive long-wavelength infrared polarimetric imaging, Appl. Opt. 53 (36) (2014) 8514–8521.
[23] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 1, IEEE. 2005, pp. 886–893.
[24] B.S. Riggan, C. Reale, N.M. Nasrabadi, Coupled auto-associative neural networks for heterogeneous face recognition, IEEE Access 3 (2015) 1620–1632.
[25] M.S. Sarfraz, R. Stiefelhagen, Deep Perceptual Mapping for Thermal to Visible Face Recognition,arXiv preprint arXiv:1507.02879. 2015.
[26] Y. He, S. Xiang, C. Kang, J. Wang, C. Pan, Cross-modal retrieval via deep and bidirectional representation learning, IEEE Trans. Multimedia 18 (7) (2016) 1363–1377.
[27] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014. pp. 1701–1708.
[28] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005., 1, IEEE. 2005, pp. 539–546.
[29] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, Advances in Neural Information Processing Systems, 2014. pp. 1988–1996.
[30] Y. Sun, D. Liang, X. Wang, X. Tang, Deepid3: Face Recognition with Very Deep Neural Networks,arXiv preprint arXiv:1502.00873. 2015.
[31] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014. pp. 1891–1898.
[32] Y. Sun, X. Wang, X. Tang, Deeply learned face representations are sparse, selective, and robust, Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2015. pp. 2892–2900.
[33] D. Chen, X. Cao, L. Wang, F. Wen, J. Sun, Bayesian face revisited: a joint formulation, Computer Vision-ECCV 2012, 2012. pp. 566–579.
[34] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: a unified embedding for face recognition and clustering, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. pp. 815–823.
[35] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, Technical Report 07-49, University of Massachusetts, Amherst. 2007.
[36] B.S. Riggan, N.J. Short, S. Hu, H. Kwon, Estimation of visible spectrum faces from polarimetric thermal faces, IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2016, IEEE. 2016, pp. 1–7.

[37] B.S. Riggan, N.J. Short, S. Hu, Thermal to Visible Synthesis of Face Images using Multiple Regions, IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018.
[38] H. Zhang, V.M. Patel, B.S. Riggan, S. Hu, Generative adversarial network-based synthesis of visible faces from polarimetric thermal faces, 2017 IEEE International Joint Conference on Biometrics (IJCB), IEEE. 2017, pp. 100–107.
[39] S. Hu, N. Short, K. Gurton, B. Riggan, Overview of polarimetric thermal imaging for biometrics, Polarization: Measurement, Analysis, and Remote Sensing XIII, vol. 10655, International Society for Optics and Photonics. 2018, pp. 1065502.
[40] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems, 2012. pp. 1097–1105.
[41] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016. pp. 770–778.
[42] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely Connected Convolutional Networks, CVPR, vol. 1, 2017. pp. 3.
[43] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in Neural Information Processing Systems, 2014. pp. 2672–2680.
[44] M. Mirza, S. Osindero, Conditional Generative Adversarial Nets,arXiv preprint arXiv:1411.1784. 2014.
[45] H. Zhang, V.M. Patel, Densely connected pyramid dehazing network, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[46] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio, The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation, IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, IEEE. 2017, pp. 1175–1183.
[47] Y. Zhu, S. Newsam, Densenet for dense flow, IEEE International Conference on Image Processing (ICIP), 2017, IEEE. 2017, pp. 790–794.
[48] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017. pp. 2881–2890.
[49] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, A. Agrawal, Context encoding for semantic segmentation, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[50] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, European Conference on Computer Vision, Springer. 2016, pp. 694–711.
[51] H. Zhang, V.M. Patel, Density-aware Single Image De-raining Using a Multi-Stream Dense Network,arXiv preprint arXiv:1802.07412. 2018.
[52] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization,arXiv preprint arXiv:1412.6980. 2014.
[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, IEEE. 2009, pp. 248–255.
[54] https://wsri.wright.edu/.
[55] X. Chen, P.J. Flynn, K.W. Bowyer, Visible-light and infrared face recognition, Proceedings of the Workshop on Multimodal User Authentication, 2003. pp. 48–55.
[56] K.A. Byrd, Preview of the newly acquired NVESD-ARL multimodal face database, Proc. SPIE, vol. 8734, 2013. pp. 34.
[57] S. Li, D. Yi, Z. Lei, S. Liao, The Casia Nir-Vis 2.0 face database, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013. pp. 348–353.
[58] S.Z. Li, Z. Lei, M. Ao, The HFB face database for heterogeneous face biometrics research, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009, IEEE. 2009, pp. 1–8.
[59] C. Reale, N.M. Nasrabadi, R. Chellappa, Coupled dictionaries for thermal to visible face recognition., ICIP, 2014. pp. 328–332.
[60] C. Reale, H. Lee, H. Kwon, Deep Heterogeneous Face Recognition Networks Based on Cross-Modal Distillation and an Equitable Distance Metric, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017. pp. 32–38.
[61] J. Lu, V.E. Liong, X. Zhou, J. Zhou, Learning compact binary face descriptor for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (10) (2015) 2041–2056.
[62] S. Saxena, J. Verbeek, Heterogeneous face recognition with CNNs, European Conference on Computer Vision, Springer. 2016, pp. 483–491.
[63] X. Liu, L. Song, X. Wu, T. Tan, Transferring deep representation for NIR-VIS heterogeneous face recognition, 2016 International Conference on Biometrics (ICB), IEEE. 2016, pp. 1–8.
[64] Z. Lei, M. Pietikäinen, S.Z. Li, Learning discriminant face descriptor, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2) (2014) 289–302.
[65] N. Yousefi, et al. Optimization of On-condition Thresholds for a System of Degrading Components with Competing Dependent Failure Processes, Reliability Engineering and System Safety (2019) https://doi.org/10.1016/j.ress.2019.106547.