

# Multi-Scale Thermal to Visible Face Verification via Attribute Guided Synthesis

Xing Di, *Student Member, IEEE* and Benjamin S. Riggan, *Member, IEEE* and Shuowen Hu, *Member, IEEE* and Nathaniel J. Short, *Member, IEEE* Vishal M. Patel, *Senior Member, IEEE*

**Abstract**—Thermal-to-visible face verification is a challenging problem due to the large domain discrepancy between the modalities. Existing approaches either attempt to synthesize visible faces from thermal faces or learn domain-invariant robust features from these modalities for cross-modal matching. In this paper, we use attributes extracted from visible images to synthesize attribute-preserved visible images from thermal imagery for cross-modal matching. A pre-trained attribute predictor network is used to extract the attributes from the visible image. Then, a novel multi-scale generator is proposed to synthesize the visible image from the thermal image guided by the extracted attributes. Finally, a pre-trained VGG-Face network is leveraged to extract features from the synthesized image and the input visible image for verification. Extensive experiments evaluated on three datasets (ARL Face Database, Visible and Thermal Paired Face Database, and Tufts Face Database) demonstrate that the proposed method achieves state-of-the-art performance. In particular, it achieves around 2.41%, 2.85% and 1.77% improvements in Equal Error Rate (EER) over the state-of-the-art methods on the ARL Face Database, Visible and Thermal Paired Face Database, and Tufts Face Database, respectively. An extended dataset (ARL Face Dataset volume III) consisting of polarimetric thermal faces of 121 subjects is also introduced in this paper. Furthermore, an ablation study is conducted to demonstrate the effectiveness of different modules in the proposed method.

**Index Terms**—Heterogeneous Face Recognition, Visual Attribute, Generative Adversarial Network.

## I. INTRODUCTION

Face Recognition (FR) is one of the most widely studied problems in computer vision and biometrics research communities due to its applications in authentication, surveillance, and security. Various methods have been developed over the last two decades that specifically attempt to address the challenges such as aging, occlusion, disguise, variations in pose, expression, and illumination. In particular, convolutional neural network (CNN) based FR methods have gained significant traction in recent years [44]. This is mainly due to the availability of large annotated datasets, affordability of graphics processing units (GPUs), and trainability of nonlinear

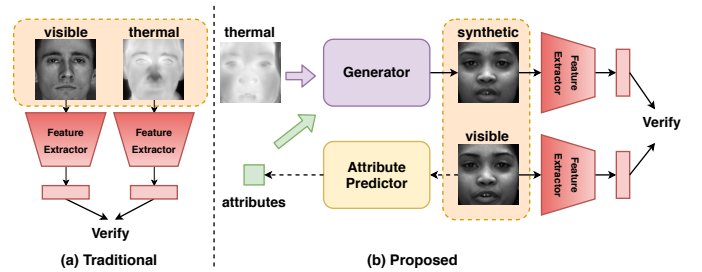


Fig. 1: (a) Traditional heterogeneous face verification approaches use the features directly extracted from different modalities for verification [21], [27], [24], [54]. (b) The proposed heterogeneous face verification approach uses a thermal face and semantic attributes to synthesize a visible face. Then, deep features extracted from the synthesized and visible faces are used for verification.

layers of deep neural networks employing activations functions (e.g., ReLU, ELU) that alleviated issues with diminishing/exploding gradients. Many deep CNN-based methods [41], [49], [56], [5], [44], [43], [8], [58] have achieved state-of-the-art performances on various FR benchmarks.

Despite the success of CNN-based methods in addressing various challenges in FR, they are fundamentally limited to recognizing face images that are collected near-infrared spectrum. In many practical scenarios such as surveillance in low-light conditions, one has to detect and recognize faces that are captured using thermal modalities [22], [47], [51], [62], [46], [27], [38], [30], [3], [2]. However, the performance of many deep learning-based methods degrades significantly when they are presented with thermal face images. For example, it was shown in [62], [46], [11], [10] that simply using deep features extracted from both thermal and visible facial images are not sufficient enough for heterogeneous face recognition. The performance degradation is mainly due to the significant distributional change between the thermal and visible domains as well as a lack of sufficient data for training the deep networks for cross-modal synthesis and matching.

Several attempts have been made to address the thermal-to-visible cross-spectrum FR problem [46], [47], [62], [11], [63]. Riggan *et al.*[47] proposed a two-step method (visible feature estimation and visible image reconstruction) to solve the heterogeneous FR problem. Zhang *et al.*[62] proposed a generative adversarial network (GAN) based method that fuses different Stokes images to synthesize a visible face image given the corresponding polarimetric thermal images. Re-

Xing Di is with the Whiting School of Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218-2608, e-mail: xing.di@jhu.edu

Benjamin S. Riggan is with the University of Nebraska, e-mail: briggan2@unl.edu

Shuowen Hu is with the U.S. Army DEVCOM Army Research Laboratory (ARL), e-mail: shuowen.hu.civ@mail.mil

Nathaniel J. Short is with Booz Allen Hamilton, e-mail: short\_nathaniel@bah.com

Vishal M. Patel is with the Whiting School of Engineering, Johns Hopkins University, e-mail: vpatel36@jhu.edu

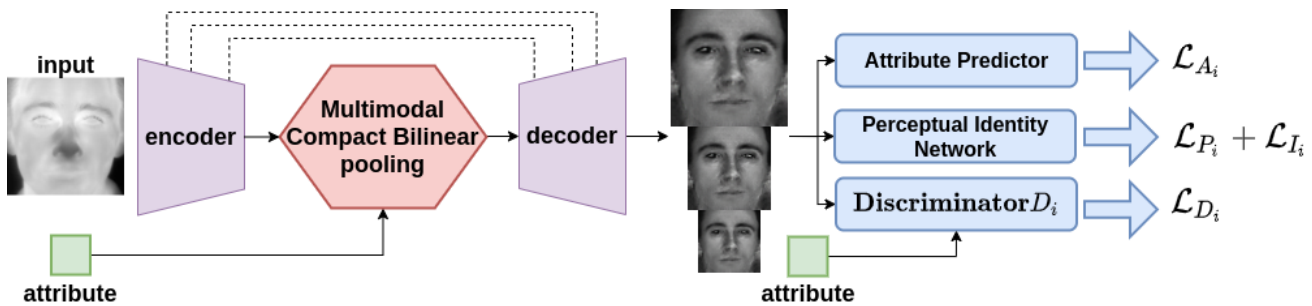


Fig. 2: A single generator with multi-scale resolution output is proposed to synthesize high-quality images by leveraging hierarchical information at different scales. Multimodal Bilinear Pooling (MCB) pooling is proposed to fuse the semantic attribute information with the image feature in the latent space. In order to make sure that the synthesized image maintains the identity and semantic attributes, a multi-purpose objective function is adopted which consists of adversarial loss  $\mathcal{L}_{D_i}$ ,  $\mathcal{L}_1$  loss, perceptual loss  $\mathcal{L}_{P_i}$ , identity loss  $\mathcal{L}_{I_i}$  and attribute preserving loss  $\mathcal{L}_{A_i}$ .

cently, Riggan *et al.*[46] developed a global and local region-based method to improve the discriminative quality of the synthesized visible imagery. Recently, Zhang *et al.*[63] introduced a multi-stream feature-level fusion method to synthesize high-quality visible images from polarimetric thermal images. Though these methods are able to synthesize photo-realistic visible face images to some extent, the synthesized results in [62], [45], [46] are still far from optimal and they tend to lose some semantic attribute information such as expression, facial hair, gender, etc. Such reconstructions may degrade the performance of thermal-to-visible face verification.

In this paper, we take a different approach to the problem of thermal-to-visible matching. Fig. 1 compares the traditional cross-modal verification problem with that of the proposed attribute-preserved heterogeneous face verification approach. Given a visible and thermal image pair, the traditional approach first extracts some features from these images and then verifies the identity based on the extracted features [27] (see Fig. 1(a)). In contrast, we propose a novel framework in which we make use of the attributes extracted from the visible image to synthesize the attribute-preserved visible image from the input thermal image for matching (see Fig. 1(b)). In particular, a pre-trained VGG-Face model [41] is used to extract the attributes from the visible image. Then, a novel Multi-Scale Attribute Preserved Generative Adversarial Network (Multi-AP-GAN) is proposed to synthesize the visible image from the thermal image guided by the extracted attributes. Finally, a pre-trained VGG-Face network is used to extract features from the synthesized and the input visible images for verification.

The proposed Multi-AP-GAN model is inspired by the recent works [29], [67], [60], [69], [68], in which deep supervision [29] is used at intermediate convolutional layers to learn better feature representations. Specifically, the Multi-AP-GAN consists of two parts: (i) a multimodal compact bilinear (MCB) pooling-based generator [13], [14], and (ii) a generator with the multi-scale architecture. The MCB pooling module fuses the given attributes with the image features. The multi-scale architecture aims to improve the synthesis image quality by leveraging hierarchical representations of CNNs at different image resolutions.

Fig. 2 provides an overview of the proposed Multi-AP-GAN

framework. A single generator with a series of distinct discriminators are employed to learn the multi-scale adversarial discrimination at different scales [55]. The generator fuses the extracted attribute vector with the image feature vector in the latent space. On the other hand, each discriminator uses triplet pairs (real image/true attributes, fake image/true attributes, fake image/wrong attributes) to not only discriminate between real and fake images but also to discriminate between the image and the attributes. In order to generate high-quality and attribute-preserved images, the generator is optimized by a multi-purpose objective function consisting of adversarial loss [15],  $L_1$  loss, perceptual loss [25], identity loss [62] and attribute preserving loss.

To summarize, this paper makes the following contributions:

- We propose a novel thermal-to-visible face verification framework in which Multi-AP-GAN is developed for synthesizing high-quality visible faces from thermal images guided by facial attributes.
- We propose a single generator with a multi-scale output architecture and a Multimodal Compact Bilinear (MCB) pooling module [13], [14] to generate high-quality visible images.
- A novel triplet-pair discriminator is proposed, where the discriminator [45] not only learns to discriminate between real/fake images as well as images/visual-attributes.
- An extended version of the ARL polarimetric thermal face database consisting of data from 121 individuals is introduced in this work.
- Extensive experiments are conducted on three different volumes of the ARL Multimodal Facial Database [22], [63] as well as the Thermal and Visible Paired Face Database [35], and comparisons are performed against several recent state-of-the-art approaches. Furthermore, an ablation study is conducted to demonstrate the improvements obtained by including semantic attribute information for synthesis.

Note that the proposed Multi-AP-GAN framework can be viewed as an extended version of our earlier paper in the 2018 BTAS proceedings [11]. However, the generators used in both papers are quite different. The generator in [11] is a single-scale generator whereas a multi-scale generator is proposed

in this paper. Furthermore, a new polarimetric thermal face dataset consisting of multimodal data from 121 subjects is introduced in this paper. Extensive experiments and analysis are presented using the new dataset as well as the Thermal and Visible Paired Face Database [35].

The rest of the paper is organized as follows. In Section II, we review a few related works on visible to thermal face synthesis and matching. Details of the proposed Multi-AP-GAN method are given in Section III. Datasets and corresponding protocols are described in Section IV. Experimental results are presented in Section V. Finally, Section VII concludes the paper with a brief summary and discussion.

## II. RELATED WORK

In this section, we review some related works on thermal-to-visible face synthesis and recognition.

### A. Feature-based Thermal-Visible Face Recognition

Traditional thermal-to-visible face verification methods first extract features from the visible and thermal images and then verify the identity based on the extracted features (See Fig. 1). Both hand-crafted and learned features have been investigated in the literature. Buddharaju *et al.* [4] proposed a method that leverages physiological information based on the superficial blood vessel network for face recognition in thermal imagery. In [57] Wesley *et al.* presented a comparative analysis of performance of automated facial expression recognition from thermal videos, visual facial videos, and their fusion using principal component analysis (PCA) based features. Gyaourova *et al.* [17] proposed a multimodal fusion method by combining information from both thermal and visible images for face recognition. Hu *et al.*[21] proposed a partial least squares (PLS) regression-based approach for heterogeneous face matching. Klare *et al.*[28] developed a generic framework for cross-modal FR based on kernel prototype nonlinear similarities. Another multiple texture descriptor fusion-based method was proposed by Bourlai *et al.* in [54] for cross-modal FR. In [24] PLS-based discriminant analysis approaches were used to correlate the thermal face images to the visible face signatures. Gurton *et al.*[16] and Nathaniel *et al.*[50], [52] proposed to use the polarization-state information of thermal emissions to enhance the performance of thermal FR. Wu *et al.*[59] introduced a disentangled variational representation for crossmodal matching in which a face representation is modeled with an intrinsic identity information and its within-person variations. He *et al.*[19] proposed a network which maps both NIR and VIS images to a compact Euclidean space for matching. Later on, they added more constraints on the representation by utilizing Wasserstein distance [20] and adversarial learning [18]. Fu *et al.* [12] proposed a framework which generates new paired images with abundant intra-class diversity to reduce the domain gap of heterogeneous face recognition.

### B. Synthesis-based Thermal-Visible Face Recognition

Synthesis-based thermal-to-visible face verification algorithms leverage the synthesized visible faces for verification.

Due to the success of CNNs and recently introduced GANs in synthesizing realistic images, various deep learning-based approaches have been proposed in the literature for thermal-to-visible face synthesis [46], [62], [66], [47], [18], [61]. For instance, Riggan *et al.*[47] proposed a two-step procedure (visible feature estimation and visible image reconstruction) to solve the cross-modal verification problem. Zhang *et al.*[62] proposed an end-to-end GAN-based approach for synthesizing photo-realistic visible face images from the corresponding polarimetric thermal images. Recently Riggan *et al.*[46] proposed a new synthesis method to enhance the discriminative quality of generated visible face images by leveraging both global and local facial regions. Zhang *et al.*[63] introduced a multi-stream fusion-based generative model for cross-modal face verification. Di *et al.*[11] proposed a GAN-based network called AP-GAN to improve the synthesized visible image by utilizing visual attributes. Di *et al.*[10] proposed another unsupervised generative model which combines features from both thermal-to-visible and visible-to-thermal synthesized images for heterogeneous face verification. Recently Pereira *et al.* [7] proposed a generic adaptation-based network for heterogeneous face recognition. He *et al.*[18] proposed a generative model for thermal-to-visible face synthesis by utilizing texture inpainting and pose correction. Another improved FusionNet was proposed in [31], which increases robustness against overfitting using dropout for a thermal-to-visible generation. This method was evaluated on the RGB-D-T dataset [39]. Recently, Mallat *et al.*[34], [6] proposed a cascaded model which is optimized by the contextual loss [36] for cross-spectrum synthesis. An attribute-guided visible face synthesis method using a conditional CycleGAN framework was proposed in [32].

## III. PROPOSED METHOD

In this section, we discuss the details of the proposed Multi-AP-GAN method (see Fig. 2). In particular, we discuss the proposed attribute predictor, multi-scale generator, a series of distinct accompanying discriminators and the loss function used to train these networks.

### A. Attribute Predictor

To efficiently extract attributes from a given visible face, an attribute predictor is fine-tuned based on the VGG-Face network [41] using the annotated attributes. This network is trained separately from Multi-AP-GAN. The fine-tuned network is used in both obtaining the visible face attributes and for capturing the attribute loss when training the generator and discriminator. When fine-tuning the network, a binary cross-entropy loss is used and the final fully-connected layer has the same dimension as the number of visual attributes. The predictor is selected based on the lowest loss error.

### B. Generator

A U-net structure [48] is used as the building block for the multi-scale generator since it is able to better capture the large receptive field and also able to efficiently address the vanishing

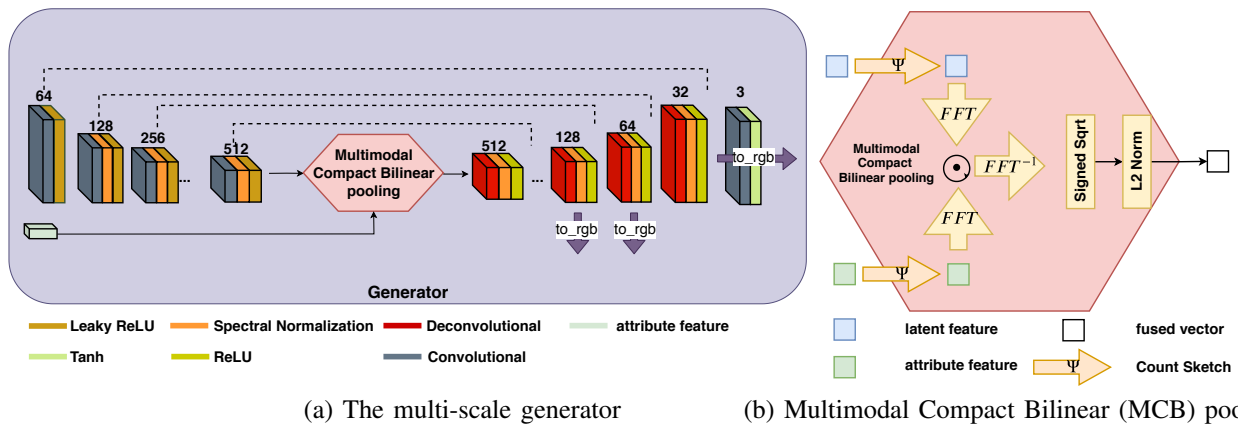


Fig. 3: The network architecture of multi-scale generator and multimodal compact bilinear (MCB) pooling in details.

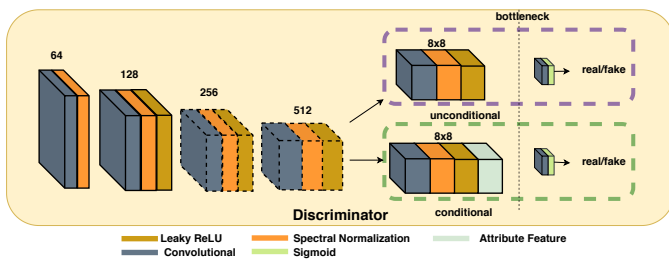


Fig. 4: An overview of the triplet-pair-input discriminator. The triplet-pair-input discriminator is composed of a conditional and an unconditional streams. The unconditional stream aims to discriminate the fake and real images. The conditional stream aims to discriminate between the image and the corresponding attributes. In order to keep the bottleneck feature map size to be consistent to  $8 \times 8$  for different input image resolution scale, a different number of downsampling layers (dash-line cubic) are utilized.

gradient problem. In addition, to effectively combine the extra facial attribute information into the building block, we fuse the attribute vector and the image feature in the latent space [45], [62], [9]. Note that the attributes are extracted from the given visible face using the fine-tuned model as discussed above. The generator architecture is illustrated in Fig 3(a).

In our experiments, we observe that simple concatenation of the two vectors (encoded image vector and attribute vector) does not work well. One possible reason is that both vectors are significantly different in terms of their dimensionality. Thus, we adopt the well-known MCB pooling method [13], [14] to overcome this issue. Instead of simple concatenation, MCB leverages the following two techniques: bilinear pooling and sketch count. Bilinear pooling is the outer-product and linearization of two vectors, where all elements of both vectors are interacting with each other in a multiplicative way. In order to overcome the high-dimension computation of bilinear pooling, Pham *et al.*[42] implemented the count sketch of the outer product of two vectors, which involves the Fast Fourier Transform ( $FFT$ ) and inverse Fast Fourier Transform ( $FFT^{-1}$ ). The architecture of the MCB module is shown in Fig 3(b).

In order to improve the quality of the synthesized visible images, the proposed single generator utilizes a multi-scale output architecture. Specifically, the generator  $G$  produces multiple outputs at different resolution scales as follows

$$G(\mathbf{x}, \mathbf{z}) = \{\hat{y}_1, \dots, \hat{y}_s\}, \quad (1)$$

where  $\mathbf{x}, \mathbf{z}$  denote the input thermal image and the extracted visual-attributes, respectively. Here,  $\{\hat{y}_1, \dots, \hat{y}_s\}$  denote the synthesized images with gradually growing resolutions and  $\hat{y}_s$  is the final output with the highest resolution  $s$ . In this work, we set  $s = 3$  where  $\hat{y}_3$  is the  $256 \times 256$  image,  $\hat{y}_2$  is the  $128 \times 128$  image, and  $\hat{y}_1$  is the  $64 \times 64$  image. These multi-scale resolution outputs act as a regularizer to the generator  $G$ . Furthermore, they shorten the error signal flow path and help to improve the training stability [68].

The multi-scale generator network, as shown in Fig. 3(a), consists of the following components: CL(64)-CNL(128)-CNL(256)-CNL(512)-CNL(512)-CNL(512)-CNL(512)-MCB(512)-DNR(512)-DNR(512)-DNR(512)-DNR(256)-DNR(128)-DNR(64)-DNR(32), where C stands for the convolutional layer (conv), L stands for LeakyReLU layer (negative\_slope=0.02), N stands for the spectral normalization layer [37], MCB indicates the Multimodal Compact Bilinear module [13], [14], D stands for the deconvolutional layer (dconv), and R corresponds to the ReLU layer. All the numbers in parenthesis indicate the channel number of the output feature maps. Table I gives the details of the generator architecture. Note that, for simplicity, spectral normalization [37], LeakyReLU and ReLU layers are omitted. In the last three layers, feature maps are converted into three-channel images by a “to\_rgb” block, which consists of one convolutional layer (parameters are indicated in parenthesis) followed by a Tanh layer.

### C. Discriminator

A series of distinct discriminators  $D_i, i = 1, \dots, s$  are utilized and trained iteratively with the generator  $G$ . For a certain discriminator at the  $i$ -th resolution scale, a patch-based discriminator [23] is leveraged and it not only aims to discriminate between real/fake images but also to discriminate between the image and the corresponding attributes. Similar to

TABLE I: Architecture details corresponding to the generator network.

	conv	conv	conv	conv	conv	conv	conv	MCB	dconv	dconv	dconv	dconv	dconv (to_rgb)	dconv (to_rgb)	dconv (to_rgb)
Input Size	256	128	64	32	16	8	4	2	2	4	8	16	32 (128)	64 (64)	128 (32)
Output Channel	64	128	256	512	512	512	512	512	512	512	512	256	128 (3)	64 (3)	32 (3)
Kernel Size	3	3	3	3	3	3	3	-	3	3	3	3	3 (3)	3 (3)	3 (3)
Stride Size	2	2	2	2	2	2	2	-	2	2	2	2	2 (1)	2 (1)	2 (1)

previous works [45], [68], [64], a triplet of paired image and attribute is given to the discriminator: *real*, *fake* and *wrong*. The *real* pair consists of a real-image ( $\mathbf{y}_i$ ) along with the corresponding true-attributes ( $\mathbf{z}$ ). The *wrong* pair consists of a real image ( $\mathbf{y}_i$ ) along with wrong attributes ( $\mathbf{z}'$ ). The *fake* pair consists of a fake-image ( $\hat{\mathbf{y}}_i$ ) with true attributes ( $\mathbf{z}$ ). The overall adversarial objective function used to train the network is as follows:

$$\mathcal{L}_G = \sum_{i=1}^s \min_G \max_{D_i} (V_{real}^i + V_{fake}^i + V_{wrong}^i),$$

$$V_{real}^i = \mathbb{E}_{\mathbf{y}_i \sim P_Y} [\log D_i(\mathbf{y}_i)] + \mathbb{E}_{\mathbf{y}_i, \mathbf{z} \sim P_{Y,Z}} [\log D_i(\mathbf{y}_i, \mathbf{z})] \quad (2)$$

$$V_{wrong}^i = \mathbb{E}_{\mathbf{y}_i, \mathbf{z}' \sim P_{Y,Z}} [\log(1 - D_i(\mathbf{y}_i, \mathbf{z}'))]$$

$$V_{fake}^i = \mathbb{E}_{\hat{\mathbf{y}}_i \sim P_{G(\mathbf{x}, \mathbf{z})}} [\log(1 - D_i(\hat{\mathbf{y}}_i))] + \mathbb{E}_{\hat{\mathbf{y}}_i \sim P_{G(\mathbf{x}, \mathbf{z})}, \mathbf{z} \sim P_Z} [\log(1 - D_i(\hat{\mathbf{y}}_i, \mathbf{z}))].$$

Specifically, each discriminator  $D_i$  has two streams: conditional stream and unconditional stream. One discriminator on  $256 \times 256$  resolution scale is illustrated in Fig. 4. The unconditional stream aims to learn the discrimination between the real and the synthesized images. This unconditional adversarial loss is back-propagated to  $G$  to make sure that the generated samples are as realistic as possible. In addition, the conditional stream aims to learn whether the given image matches the given attributes or not. This conditional adversarial loss is back-propagated to  $G$  so that it generates samples that are attribute-preserving.

Fig. 4 gives an overview of a discriminator at  $256 \times 256$  resolution scale. This discriminator consists of 6 convolutional blocks for both conditional and unconditional streams. Details of these convolutional blocks are as follows:

CL(64)-CNL(128)-CNL(256)-CNL(512)-C<sup>†</sup>NL(512)-C<sup>†</sup>S(1), where S stands for the Sigmoid activation layer. Note that the only difference between the unconditional and conditional stream is the concatenation of the attribute vector at the fifth convolutional block. For different discriminator,  $D_i$  at different resolution scale, the number of convolutional down-sample blocks (blocks with dotted lines in Fig. 4) vary, but we keep the bottleneck feature map at the same size (i.e.  $8 \times 8$ ). The architecture details corresponding to the other discriminators are given in Table II.

#### D. Loss Function

The generator is optimized by minimizing the following loss

$$\mathcal{L}_{Multi-AP-GAN} = \mathcal{L}_G + \lambda_A \mathcal{L}_A + \lambda_P \mathcal{L}_P + \lambda_I \mathcal{L}_I + \lambda_1 \mathcal{L}_1, \quad (3)$$

<sup>†</sup>unconditional and conditional streams are shortened for brevity.

TABLE II: Architecture details corresponding to different discriminators. Numbers in parenthesis indicate the channel number of the output feature maps. The convolutional layers have stride size 2.

Discriminator 64x64	Discriminator 128x128	Discriminator 256x256
Convolutional (64) LeakyReLU	Convolutional (64) LeakyReLU	Convolutional (64) LeakyReLU
Convolutional (128) Spectral Norm LeakyReLU	Convolutional (128) Spectral Norm LeakyReLU	Convolutional (128) Spectral Norm LeakyReLU
Convolutional <sup>†</sup> (256) Spectral Norm LeakyReLU	Convolutional (256) Spectral Norm LeakyReLU	Convolutional (256) Spectral Norm LeakyReLU
Convolutional <sup>†</sup> (1) Sigmoid	Convolutional <sup>†</sup> (512) Spectral Norm LeakyReLU	Convolutional (512) Spectral Norm LeakyReLU
	Convolutional <sup>†</sup> (1) Sigmoid	Convolutional <sup>†</sup> (512) Spectral Norm LeakyReLU
		Convolutional <sup>†</sup> (1) Sigmoid

where  $\mathcal{L}_G$  is the multi-scale adversarial loss in Eq (2),  $\mathcal{L}_P$  is the perceptual loss,  $\mathcal{L}_I$  is the identity loss,  $\mathcal{L}_A$  is the attribute loss,  $\mathcal{L}_1$  is the loss based on the  $L_1$ -norm between the target and the reconstructed image, and  $\lambda_P, \lambda_I, \lambda_A, \lambda_1$  are the corresponding weights.

1) *Multi-scale Perceptual and Identity Loss*: Perceptual loss was originally introduced by Johnson *et al.*[25] for style transfer and super-resolution. It has been observed that the perceptual loss produces visually pleasing results than  $L_1$  or  $L_2$  loss. The perceptual and identity losses are defined as follows

$$\mathcal{L}_{P,I} = \sum_{i=1}^s \sum_{c=1}^3 \sum_{w=1}^W \sum_{h=1}^H \|F(\hat{\mathbf{y}}_i)^{c,w,h} - F(\mathbf{y}_i)^{c,w,h}\|_1, \quad (4)$$

where  $F$  represents a non-linear CNN feature. VGG-16 [53] is used to extract features in this work.  $C, W, H$  are the dimensions of features from a certain level of the VGG-16, which are different for perceptual and identity losses. Since the deeper convolutional layer captures more semantic information, we choose deeper convolutional feature maps as the identity loss.

In addition, multi-scale  $L_1$  loss between the synthesized image  $\hat{\mathbf{y}}_i$  and the corresponding real image  $\mathbf{y}_i$  is used to capture the low-frequency information, which is defined as follows

$$\mathcal{L}_1 = \sum_{i=1}^s \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_1. \quad (5)$$

2) *Multi-scale Attribute Loss*: Inspired by the perceptual loss, we define an attribute preserving loss, which measures the

error between the attributes of the synthesized image and the real image. To make sure the pre-trained model captures the facial attribute information, we fine-tune the pre-trained VGG-Face network on the annotated attribute dataset and regard the fine-tuned attribute classifier as the pre-trained model for the attribute preserving loss. Similar to the perceptual loss, the  $\mathcal{L}_A$  is defined as follows

$$\mathcal{L}_A = \sum_{i=1}^s \|Q(\hat{y}_i) - Q(y_i)\|_1, \quad (6)$$

where  $Q$  is the fine-tuned attribute predictor network. The output vectors are from the last layer. As a result the feature dimensions  $C, W, H$  are omitted in (6). By feeding such attribute information into the generator during training, the generator  $G$  is able to learn semantic information corresponding to the face.

### E. Implementation

The entire network is trained in Pytorch on a single Nvidia Titan-X GPU. During the Multi-AP-GAN training, the  $L_1$ , perceptual and identity loss parameters are chosen as  $\lambda_1 = 10$ ,  $\lambda_P = 2.5$ ,  $\lambda_I = 0.5$ , respectively. The ADAM [26] is implemented as the optimization algorithm with parameter  $\beta_{t_1} = (0.5, 0.999)$  and batch size is set equal to 1. The total epochs are 200. For the first 100 epochs, we fix the learning rate as 0.0002 and for the remaining 100 epochs, the learning rate was decreased by 1/100 after each epoch. The feature maps for the perceptual and the identity loss are from the relu1-1 and the relu2-2 layers, respectively. In order to fine-tune the attribute predictor network, we manually annotate images with the attributes tabulated in Table III.

TABLE III: The facial attributes used in this work.

attributes	Arched_Eyebrows, Big_Lips, Big_Nose, Bushy_Eyebrows, Male, Mustache, Narrow_Eyes, No_Beard, Mouth_Slightly_Open, Young
------------	--

## IV. DATASETS AND PROTOCOLS

In this section, we describe the datasets and the protocols that we use to conduct experiments. In particular, we describe the new extended ARL Polarimetric thermal face dataset and the corresponding protocol that we use in this paper.

### A. Extended Polarimetric Thermal Face Dataset

In many recent approaches, the polarization-state information of thermal emissions has been used to achieve improved cross-spectrum face recognition performance [22], [47], [51], [62], [46] since it captures geometric and textural details of faces that are not present in the conventional thermal facial images [51], [22]. A polarimetric thermal image consists of three Stokes images:  $S_0, S_1, S_2$  where  $S_0$  indicates the conventional total intensity thermal image,  $S_1$  captures the horizontal and vertical polarization-state information,  $S_2$  captures the diagonal polarization-state information [22]. Similar to [62], [46], we also refer to Polar as the three channel

polarimetric image concatenated with  $S_0, S_1$  and  $S_2$ . These Stokes images along with the visible and the polarimetric images corresponding to a subject in the ARL dataset [22] are shown in Fig. 5. It can be observed that  $S_1, S_2$  tend to preserve more textural details compared to  $S_0$ .

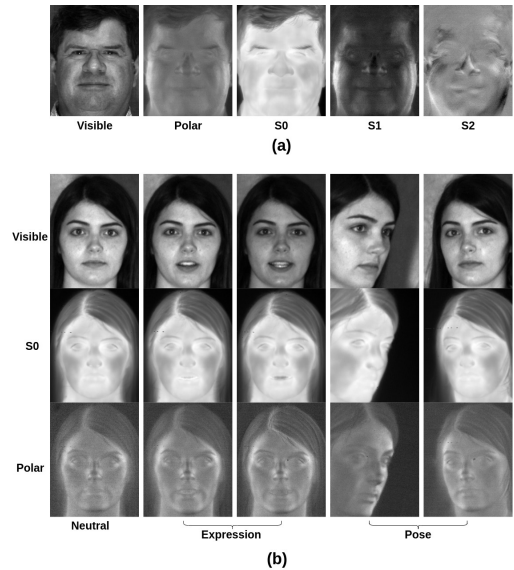


Fig. 5: Sample images from the ARL dataset. (a) Visible, polarimetric thermal, and Stokes images ( $S_0, S_1, S_2$ ) corresponding to a subject from the ARL dataset [22]. (b) Sample visible, conventional thermal and polarimetric thermal images with different variations from the ARL Dataset Volume III.

The U.S. Army DEVCOM Army Research Laboratory (ARL) multimodal face dataset consists of polarimetric thermal and visible face image pairs in three volumes. Volume I consists of the polarimetric thermal and visible images from 60 subjects, which were collected by the U.S. Army Research Laboratory in 2014-2015. Frontal imagery with different ranges and expressions are included. Details regarding this volume can be found in [22] and [63]. Volume II consists of images from 51 subjects collected at a Department of Homeland Security test facility. As described in [63], while the participants of the Volume I subset consisted exclusively of the ARL employees, the participants of the Volume II collect were recruited from the local community in Maryland, resulting in more demographic diversity. In addition, frontal imagery with various expressions is included in this volume.

In this paper, we present an extension of the dataset which was collected by ARL across 11 different sessions over 6 days. We refer to this extended dataset as Volume III hereinafter. Volume III contains polarimetric thermal and visible facial signatures from 121 subjects collected at Johns Hopkins University Applied Physics Laboratory as part of an IARPA government testing event. There are a total of 5419 polarimetric thermal and visible image pairs with significant variations (Fig. 5) such as expression, off-pose, glasses, etc. These variations make the dataset more challenging for cross-modal face verification. Note that this extended database is available upon request.

To be consistent with previous methods [63], [22], the experimental protocols are defined as follows:

**Protocol I:** The Protocol I is evaluated on Volume I, which consists of frontal imagery with range and expression variations (including neutral expression). Images from 30 subjects with eight samples for each subject are used as the training split. Images from the other 30 subjects with eight samples for each subject are used as the test split. All the samples in training and test split are randomly chosen from 60 subjects. Results are evaluated on five random splits. Note that there are no overlapping subjects between training and test splits.

**Protocol II:** The Protocol II is evaluated on the extended 111 subject dataset which contains the images from both Volume I and Volume II. In particular, 85-subject images are used as the training split and the other 26-subject images are denoted as the test split. The 85-subject images in training split consist of all 60-subject images in Volume I and another 25-subject images randomly selected from Volume II. The other 26-subject images in Volume II are selected as the test split. As before, results are evaluated on five random splits [63]. Note that Volume II consists of frontal imagery with expression variations only (including neutral expression).

**Protocol III:** The Protocol III is evaluated only on the Volume III data consisting of images from 121 subjects. Volume III includes frontal and off-pose imagery (excludes extreme pose, e.g. profile), and expression variation (including neutral expression). Images from 96 randomly chosen subjects are used as the training split and the images from the remaining 25 subjects are used as the test split. Results are evaluated on five random splits.

### B. Visible and Thermal Paired Face Database

In addition to the ARL dataset, the proposed method is evaluated on a recently introduced Visible and Thermal Paired Face Database [35]. This dataset contains thermal and visible image pairs corresponding to 50 subjects. Each subject participated in two different sessions separated by a time interval of 3 to 4 months. This dataset includes 21 face images per subject in each session. These images correspond to different facial variations in illumination, head pose, expression and occlusion. In total, 4200 images are included in this dataset.

**Protocol:** Images corresponding to randomly chosen 30 subjects are used as the training split and the images from the remaining 20 subjects are used as the test split. This results in 630 paired training images and 420 paired testing images. There is no overlap among subjects in the training and the test sets. Results are evaluated on five random splits.

### C. Tufts Face Database

We also evaluate the proposed method on a recently proposed Tufts Face Database [40], which contains 1532 paired visible and thermal face images from 112 subjects. For each subject multiple images are taken in different conditions. In particular, each subject has images in 9 different poses, 4 expressions and 1 occlusion with eye glasses. Sample images from this dataset are shown in 6. The Tufts dataset [40] is

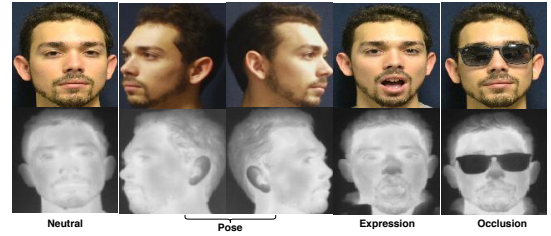


Fig. 6: Sample thermal and visible images from the Tufts Face Database [40] with different variations.

more difficult than the other two datasets as it contains less number of images per person in each variation.

**Protocol:** Similar to the previous protocols, images corresponding to 90 subjects are used for training and the images from the remaining 22 subjects are used for testing. This results in about 1232 paired data for training and 300 paired data for testing. There is no overlap among subjects in the training and the test sets. Results are reported based the evaluations on five random splits.

### D. Preprocessing

In addition to the standard preprocessing, two more preprocessing steps are used for the proposed method. First, the faces in the visible images are detected by MTCNN [65]. Then, a standard central crop method is used to crop the detected faces. Since MTCNN is implementable on the visible images only, we use the same detected rectangle coordinations to crop the thermal images, which were already aligned to the same canonical coordinates as the visible images. After preprocessing, all the images are scaled and saved as  $256 \times 256$  16-bit PNG files.

### E. Metrics

Once the visible image is synthesized from the input probe thermal image, we use a pre-trained VGG-Face model [41] to extract features from the synthesized visible probe image as well as the visible gallery image to perform cross-modal face verification. In particular, the verification score is calculated using the cosine similarity between the two feature vectors. The cross-modal verification performance of different methods is evaluated using the Receiver Operating Characteristic (ROC) curve, Area Under the Curve (AUC) and Equal Error Rate (EER) measures.

## V. EXPERIMENTAL RESULTS

In this section, we demonstrate the effectiveness of the proposed approach by conducting various experiments on the datasets described in the previous section. Since the ARL Dataset contains both conventional thermal ( $S_0$ ) and polarimetric thermal modalities, we conduct the following two cross-modal face verification experiments on the ARL dataset: 1) Conventional thermal ( $S_0$ ) to Visible (Vis) and 2) Polarimetric thermal (Polar) to Visible (Vis). On the other hand, the Visible and Thermal Paired Face Database and the Tufts Face Database do not contain polarimetric thermal

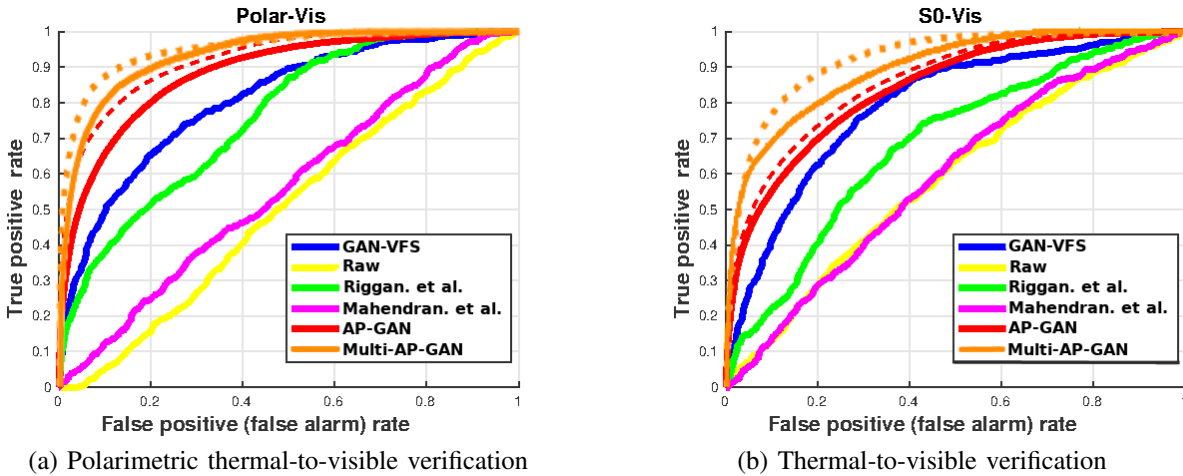


Fig. 7: The ROC curve comparison on Protocol I with several state-of-the-art methods: (a) Polarimetric thermal-to-visible verification performance. (b) S0-to-Visible verification performance. Note that the dotted lines indicate results based on the ground-truth attributes. The gap between the results with ground-truth attributes and that with predicted attributes demonstrate the degradation caused by the attribute predictor.

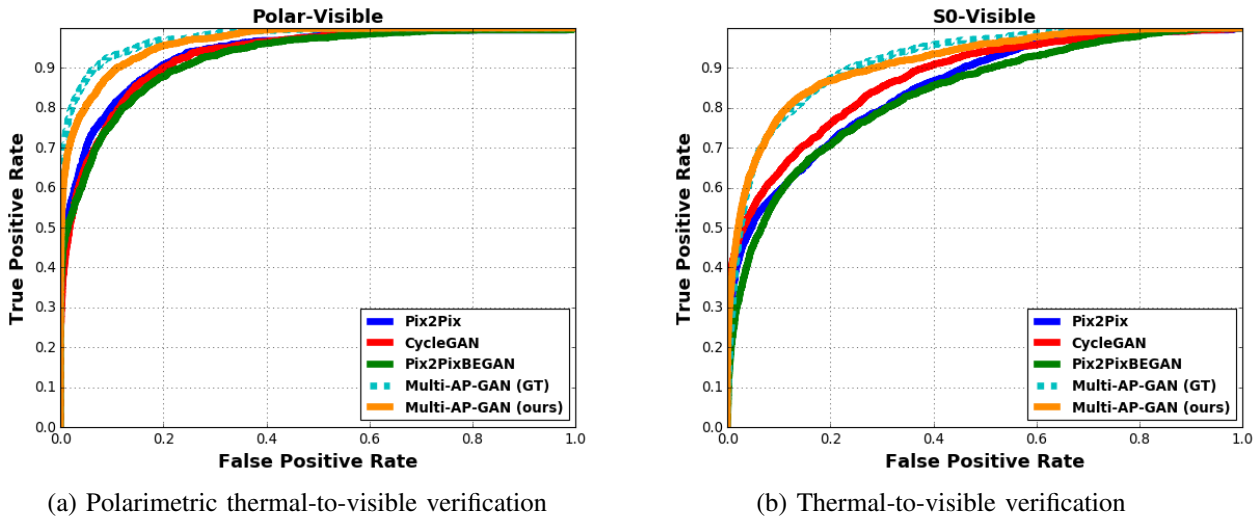


Fig. 8: The ROC curve comparison on Protocol II with several state-of-the-art methods: (a) Polarimetric thermal-to-visible verification performance. (b) S0-to-Visible verification performance. Note that the dotted lines indicate results based on the ground-truth attributes. The gap between the results with ground-truth attributes and that with predicted attributes demonstrate the degradation caused by the attribute predictor.

images. As a result, we only conduct thermal-to-visible cross-domain face verification experiments on these datasets.

We evaluate and compare the performance of the proposed method with that of the following recent state-of-the-art methods [62], [33], [47], [46], [11], [63], [34], [6]. Note that our previous work [11] can be viewed as a single scale version of the proposed method. In particular, in [11], we synthesize images at a particular scale which has the same resolution as the input. We also conduct experiments with another baseline method called, Multi-AP-GAN (GT), where we use the ground-truth attributes in our method rather than automatically predicting them using the proposed attribute predictor. This baseline will clearly determine how effective the proposed attribute predictor is in determining the attributes

from unconstrained visible faces.

#### A. Results on the ARL Face Dataset

Fig. 7 shows the performance corresponding to Protocol I on two different experimental settings (i.e S0-to-visible and Polar-to-visible). Compared with other state-of-the-art methods in Fig. 7, the proposed method performs better with a larger AUC and lower EER scores. In addition, it can be observed that the performance corresponding to the Polar modality is better than the S0 modality, which also demonstrates the advantage of using the polarimetric thermal images than the conventional thermal images. In addition, the gap between the results with ground-truth attributes (dash-line) and that with the predicted attributes (solid-line) demonstrates the degradation caused by



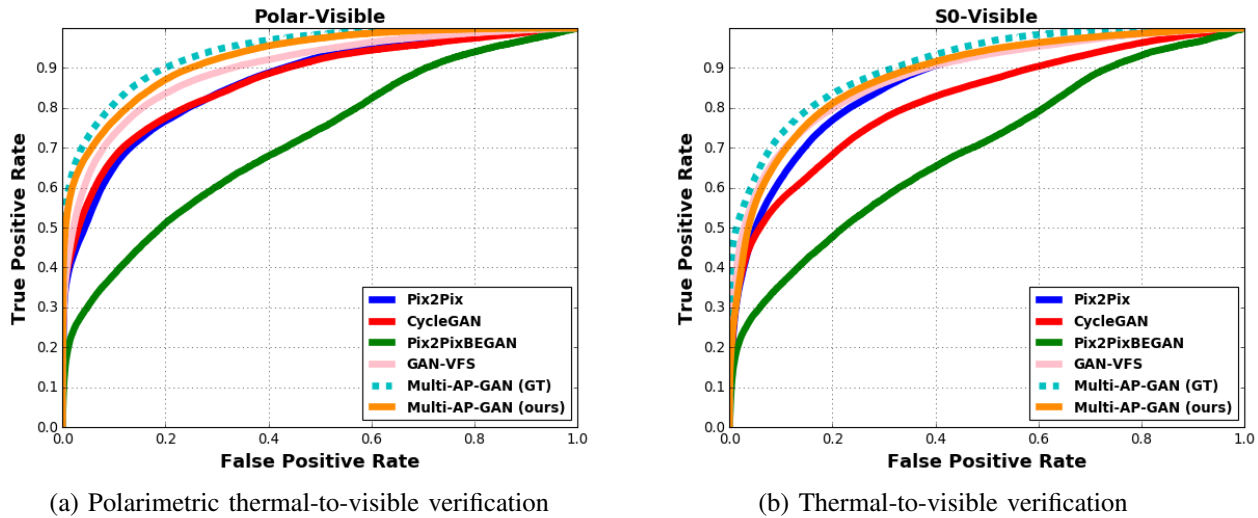


Fig. 9: The ROC curve comparison on Protocol III with several state-of-the-art methods: (a) Polarimetric thermal-to-visible verification performance. (b) S0-to-Visible verification performance. Note that the dotted lines indicate results based on the ground-truth attributes. The gap between the results with ground-truth attributes and that with predicted attributes demonstrate the degradation caused by the attribute predictor.

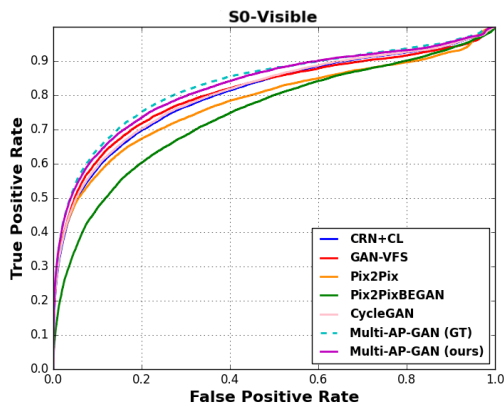


Fig. 10: The ROC curve comparison on Thermal-Visible Paired Database [35]. Note that the dotted lines indicate results based on the ground-truth attributes. Similarly, the gap between the results with ground-truth attributes and that with predicted attributes demonstrate the degradation caused by the attribute predictor.

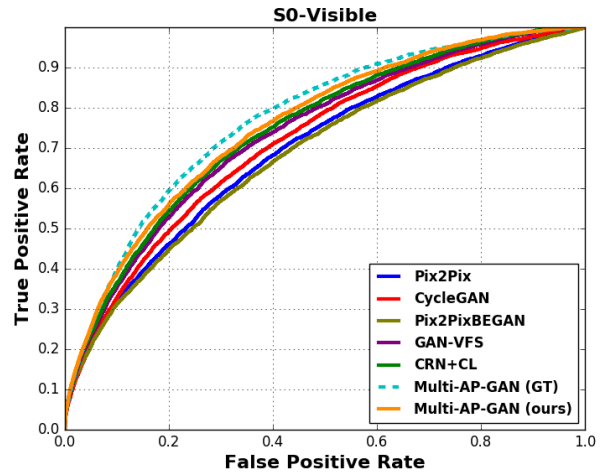


Fig. 11: The ROC curve comparison on the Tufts Face Database [40]. Note that the dotted lines indicate results based on the ground-truth attributes. Similarly, the gap between the results with the ground-truth attributes and that with predicted attributes demonstrate the degradation caused by the attribute predictor.

the attribute predictor. The quantitative comparisons, as shown in the Table IV, also demonstrate the effectiveness of the proposed method. In addition, compared with the previous single scale resolution method [11], the proposed multi-scale algorithm achieves significant improvement: around 4% and 6% on the conventional and polarimetric thermal modalities, respectively. These improvements demonstrate the effectiveness of the proposed multi-scale synthesis algorithm.

Furthermore, we also show some visual comparisons in Fig. 12. The first row in Fig. 12 shows one synthesized sample using S0. The second row shows the same synthesized sample using a polarimetric thermal image. It can be observed that the results of Riggan *et al.*[47] do capture the overall face

structure but it tends to lose some facial details. Results of Mahendran *et al.*[33] are poor compared to [47]. Results of Zhang *et al.*[62] are more photo-realistic but tend to lose some attribute information. The proposed Multi-AP-GAN not only generates photo-realistic images but also preserves attributes on the reconstructed images.

Fig. 8 and Table V show the performance of different methods on Protocol II. These results also demonstrate the superiority of the proposed method. Note that the performance of many methods is slightly better in Protocol II than Protocol I. This is mainly due to the fact that the training dataset is

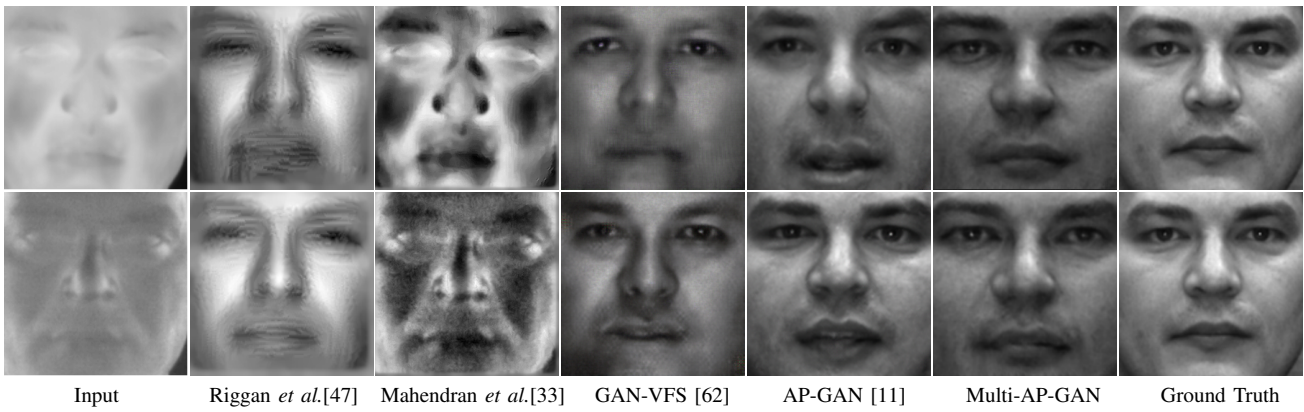


Fig. 12: The visual comparison of synthesized samples from different methods: Riggan *et al.*[47], Mahendran *et al.*[33], GAN-VFS [62], AP-GAN [11], Multi-AP-GAN, Ground Truth. The first row results correspond to the S0 image, and the second row results correspond to the Polar image.

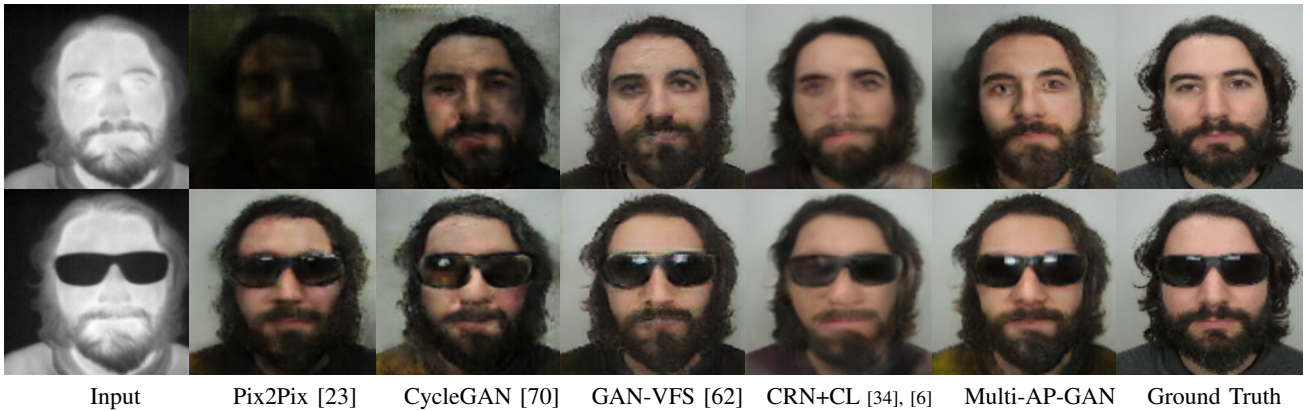


Fig. 13: The visual comparison of synthesized images corresponding to Pix2Pix[23], CycleGAN [70], GAN-VFS [62], CRN+CL [34], [6], Multi-AP-GAN (ours) from the Visible and Thermal Paired Face Database [35].



Fig. 14: Some failure cases. Note that extreme pose, illumination and occlusion variations cause the proposed method to synthesize poor quality images.

larger in Protocol II than Protocol I.

Protocol III results corresponding to different methods are shown in Fig. 9 and Table VI. Note that face images in this volume include many variations such as expression, pose, illuminations and occlusion (glasses). As a result, the performance of the methods compared is slightly lower than what we observed in Protocol I and Protocol II. In general, the proposed method performs favorably against the state-of-the-art methods. Note that Pix2PixBEGAN method [23], [1] fails to generate good quality visible faces from profile thermal face images. As a result, Pix2PixBEGAN method performs poorly on this dataset.

We further analyze the cross-modal verification performance of different methods on different variation settings on Protocol III. The corresponding results are shown in Table VII. Since variations like occlusion and illumination are not included in some subjects, we only use three variations (neutral, expression, and pose) which are included in all subjects. As can be seen from Table VII, the performance degradation mainly comes from pose variations.

### B. Results on the Visible and Thermal Paired Face Database

Table VIII shows the performance of different methods on the Visible and Thermal Paired Face Database. Compared to

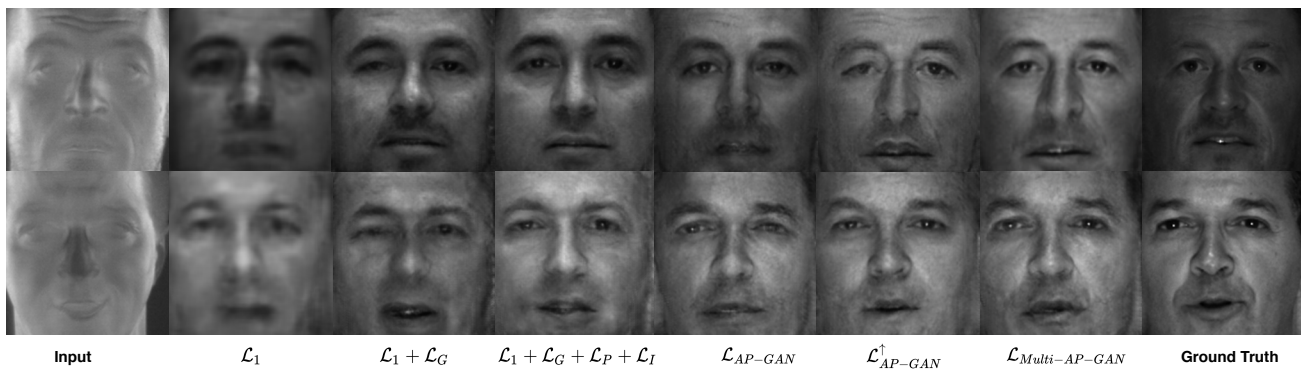


Fig. 15: The visual results of the ablation study for different experimental settings. Given input polarimetric thermal image, synthesized results using different combination of losses and resolutions are shown successively from left to right. One intermediate synthesis results  $\mathcal{L}_{AP-GAN}^{\uparrow}$ , which utilizes 2-scale resolutions  $128 \times 128$  and  $256 \times 256$ , is shown here to demonstrate the progressive improvements obtained by adding multi-scale.

the ARL Face dataset, the performance of every method is lower on this dataset. This is mainly due to the fact that this dataset is small in size and contains many facial variations. In general, the proposed method performs favorably against the previous methods.

In addition, following the analysis presented in [35], we also analyze how different variations (i.e. illumination, pose, expression, occlusion) influence the cross-spectrum matching performance of our method. As can be seen from the results in Table IX illumination and pose variations are the two variations that affect the performance of our method the most. This analysis is based on the proposed method implemented with the ground-truth visual attributes.

We also show some visual results in Fig. 13. It can be observed that Pix2Pix [23] and CycleGAN [70] methods generate poor quality images with many artifacts. GAN-VFS *et al.*[62] is able to synthesize better quality images. However, this method also introduces some artifacts around the eyes and mouth regions. The proposed Multi-AP-GAN method not only generates photo-realistic images but also preserves attributes on the synthesized images. We also show some images in Fig. 14 in which the proposed method is not able to produce good quality images. From these images we see that extreme pose, occlusion and illumination variations cause the proposed method to produce poor quality images.

### C. Results on the Tufts Face Database

Table X and Fig. 11 show the performance of different methods on the Tufts Face Dataset [40]. Compared to the previous two datasets, this dataset is more challenging due to a large number of pose and expression variations as well as a few number of images per variation, which leads to the lower performance of every method. In general, our method outperforms the other baseline methods on this challenging dataset by improvements on 1.8 % EER and 2.4% AUC scores respectively.

### D. Ablation Study

In order to demonstrate the effectiveness of different modules in the proposed method, we conduct the following

ablation study using the Polarimetric thermal modality in the ARL dataset on Protocol I:

- 1) Polar to Visible estimation with only  $\mathcal{L}_1$  (as defined in Eq. (5))
- 2) Polar to Visible estimation with  $\mathcal{L}_1$  and  $\mathcal{L}_G$  (as defined in Eq. (2))
- 3) Polar to Visible estimation with  $\mathcal{L}_1$ ,  $\mathcal{L}_G$ , perceptual loss  $\mathcal{L}_P$  and identity loss  $\mathcal{L}_I$ , which are defined as in Eq. (4).
- 4) Polar to Visible estimation with all the losses as defined in Eq. (3), by utilizing various solution scales:  $\mathcal{L}_{AP-GAN}$  ( $256^2$ ),  $\mathcal{L}_{AP-GAN}^{\uparrow}$  ( $128^2, 256^2$ ),  $\mathcal{L}_{Multi-AP-GAN}$  ( $64^2, 128^2, 256^2$ ) respectively.

Fig. 17 shows the ROC curves corresponding to each experimental setting. From this figure, we can observe that using all the losses together as  $\mathcal{L}_{Multi-AP-GAN}$  can obtain the best performance. Compared to the results between  $\mathcal{L}_1$  and  $\mathcal{L}_1 + \mathcal{L}_G$ , we can observe the enhancement provided by adding the adversarial loss. Compared with the results between  $\mathcal{L}_1 + \mathcal{L}_G$  and  $\mathcal{L}_1 + \mathcal{L}_G + \mathcal{L}_P + \mathcal{L}_I$ , we can observe the improvements obtained by adding the perceptual and identity losses. On the other hand, one can clearly see the significance of fusing the semantic attribute information with the image feature in the latent space by comparing the results between  $\mathcal{L}_1 + \mathcal{L}_G + \mathcal{L}_P + \mathcal{L}_I$  and  $\mathcal{L}_{AP-GAN}$ . Additionally, looking at the comparison with  $\mathcal{L}_{AP-GAN}$ ,  $\mathcal{L}_{AP-GAN}^{\uparrow}$  and  $\mathcal{L}_{Multi-AP-GAN}$ , one can see the successive improvements by leveraging the multi-scale information.

Besides the ROC curves, we also show the visual results for each experimental setting in Fig. 15. Given the input Polar image, the synthesized results from different experimental settings are shown in Fig. 15. It can be observed that  $\mathcal{L}_1$  captures the low-frequency features of images very well.  $\mathcal{L}_1 + \mathcal{L}_G$  can capture both low-frequency and high-frequency features in the image. However, it adversely introduced distortions and artifacts in the synthesized image. In addition, optimizing  $\mathcal{L}_P + \mathcal{L}_I$  suppresses these distortions to some extent. Finally, fusing attributes into the loss on with leveraging multi-scale resolution (i.e.  $\mathcal{L}_{Multi-AP-GAN}$ ) can not only improving the

TABLE IV: ARL Protocol I verification performance comparisons among the baseline methods, state-of-the-art methods, and the proposed Multi-AP-GAN method for both polarimetric thermal (Polar) and conventional thermal (S0) cases.

Method	AUC(Polar)	AUC(S0)	EER(Polar)	EER(S0)
Raw	50.35%	58.64%	48.96%	43.96%
Mahendran <i>et al.</i> [33]	58.38%	59.25%	44.56%	43.56%
Riggan <i>et al.</i> [47]	75.83%	68.52%	33.20%	34.36%
GAN-VFS <i>et al.</i> [62]	79.90%	79.30%	25.17%	27.34%
Riggan <i>et al.</i> [46]	85.43%	82.49%	21.46%	26.25%
AP-GAN [11]	88.93% ± 1.54%	84.16% ± 1.54%	19.02% ± 1.69%	23.90% ± 1.52%
AP-GAN (GT) [11]	91.28% ± 1.68%	86.08% ± 2.68%	17.58% ± 2.36%	23.13% ± 3.02%
Multi-stream GAN [63]	<b>96.03%</b>	85.74%	11.78%	23.18%
Multi-AP-GAN (ours)	93.61% ± 1.46%	<b>90.14% ± 2.17%</b>	14.24% ± 1.91%	<b>18.20% ± 2.65%</b>
Multi-AP-GAN (GT) (ours)	95.29% ± 1.39%	<b>92.72% ± 2.03%</b>	<b>11.22% ± 1.89%</b>	<b>16.05% ± 2.15%</b>

TABLE V: ARL Protocol II verification performance comparisons among the baseline methods and the proposed method for both polarimetric thermal (Polar) and conventional thermal (S0) cases.

Method	AUC (Polar)	AUC(S0)	EER(Polar)	EER(S0)
Raw	66.85%	63.66%	37.85%	40.93%
Pix2Pix [23]	93.66% ± 1.07%	85.09% ± 1.48%	13.73% ± 1.38%	23.12% ± 1.14%
Pix2PixBEGAN [23], [1]	92.16% ± 1.09%	83.69% ± 1.28%	15.38% ± 1.45%	26.22% ± 1.16%
CycleGAN [70] (supervised)	93.11% ± 1.02%	87.29% ± 1.13%	15.19% ± 1.02%	20.99% ± 1.19%
Multi-stream GAN [63]	<b>98.00%</b>	–	7.99%	–
Multi-AP-GAN (ours)	96.55% ± 1.12%	<b>91.43% ± 0.93%</b>	10.17% ± 1.01%	<b>15.86% ± 2.13%</b>
Multi-AP-GAN (GT) (ours)	97.68% ± 0.78%	<b>91.88% ± 0.87%</b>	<b>7.69% ± 1.39%</b>	<b>15.29% ± 2.36%</b>

TABLE VI: ARL Protocol III verification performance comparisons among the baseline methods and the proposed method for both polarimetric thermal (Polar) and conventional thermal (S0) cases.

Method	AUC (Polar)	AUC(S0)	EER(Polar)	EER(S0)
Raw	73.43%	76.71%	33.56%	30.76%
Pix2Pix [23]	86.78% ± 1.84%	86.65% ± 1.48%	21.92% ± 1.26%	23.12% ± 1.77%
Pix2PixBEGAN [23], [1]	71.29% ± 1.88%	69.42% ± 1.84%	33.83% ± 1.68%	36.88% ± 1.76%
CycleGAN [70] (supervised)	86.77% ± 1.77%	81.80% ± 1.67%	21.48% ± 1.11%	25.86% ± 1.36%
GAN-VFS ‡ [62]	90.20% ± 1.85%	87.10% ± 1.52%	18.53% ± 1.21%	20.22% ± 1.92%
Multi-AP-GAN (ours)	<b>92.29% ± 1.48%</b>	<b>88.49% ± 1.87%</b>	<b>16.26% ± 1.12%</b>	<b>19.25% ± 1.62%</b>
Multi-AP-GAN (GT) (ours)	<b>93.72% ± 1.08%</b>	<b>90.99% ± 1.13%</b>	<b>14.75% ± 1.36%</b>	<b>17.81% ± 1.63%</b>

TABLE VII: Protocol III verification performance with respect to different variations.

Variations	AUC (Polar)	AUC(S0)	EER(Polar)	EER(S0)
Neutral	96.77% ± 1.25%	94.69% ± 1.17%	12.50% ± 2.09%	13.38% ± 1.48%
Expression	96.77% ± 1.91%	92.38% ± 1.40%	10.05% ± 2.02%	15.18% ± 1.58%
Pose	86.62% ± 2.39%	82.35% ± 2.54%	22.45% ± 1.84%	25.76% ± 1.95%
Average	93.72% ± 1.08%	90.99% ± 1.13%	14.75% ± 1.36%	17.81% ± 1.63%

TABLE VIII: Visible and Thermal Paired Face Database verification performance comparisons among the baseline methods and the proposed method for the conventional thermal case.

Method	AUC	EER
Raw	69.54%	35.39%
Pix2Pix [23]	78.66% ± 1.48%	28.39% ± 1.14%
Pix2PixBEGAN [23], [1]	73.69% ± 1.82%	34.22% ± 1.61%
CycleGAN [70] (supervised)	80.24% ± 1.31%	26.72% ± 1.39%
GAN-VFS ‡ [62]	80.44% ± 1.03%	26.33% ± 1.19%
CRN + CL ‡ [34], [6]	81.25% ± 1.01%	26.01% ± 1.23%
Multi-AP-GAN (ours)	<b>81.73% ± 0.93%</b>	<b>25.68% ± 1.56%</b>
Multi-AP-GAN (GT) (ours)	<b>82.68% ± 0.87%</b>	<b>23.16% ± 0.98%</b>

performance but also preserves facial attributes. In our study, we do not see significant more improvement by utilizing more than 3-scale resolutions.

In addition, we analyze the effect of attributes on the synthesized images in Figure 16. In particular, given the input gallery image, we examine how attributes help in synthesizing a visible image from a thermal probe image. If the probe image and the input gallery image share the same identity then Multi-

TABLE IX: Verification performance with respect to different variations on the Visible and Thermal Paired Face Database.

Variations	AUC	EER
Illumination	73.35% ± 0.25%	32.60% ± 0.43%
Expression	97.25% ± 0.68%	7.45% ± 1.74%
Pose	78.25% ± 1.03%	28.75% ± 0.93%
Occlusion	83.98% ± 1.33%	24.02% ± 1.06%
Average	82.68% ± 0.87%	23.16% ± 0.98%

AP-GAN is able to generate attribute preserving visible image. On the other hand, if the probe image’s identity is different from that of the gallery image then the proposed method is not able to synthesize identity preserving visible face. However, the attributes are still preserved on the synthesized image. This analysis further demonstrates that the proposed Multi-AP-GAN method learns the cross-spectral (thermal-to-visible)

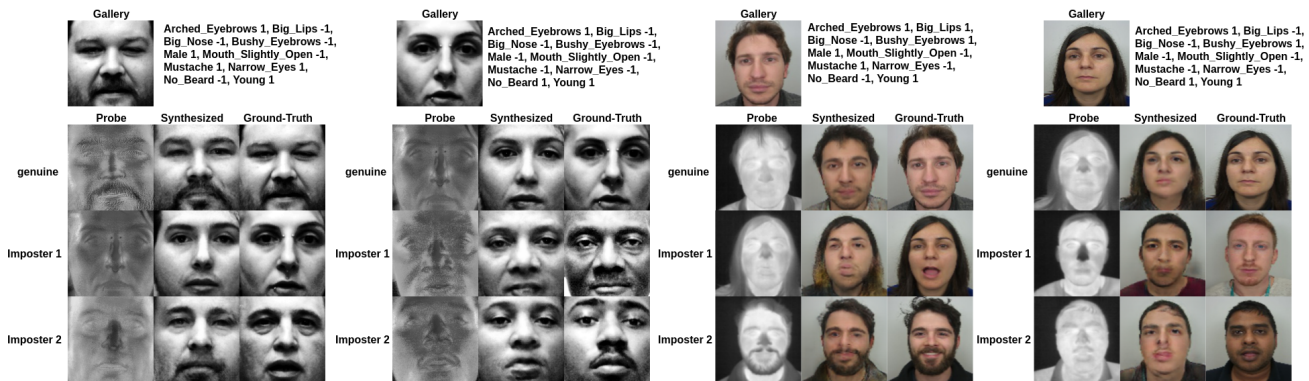


Fig. 16: Analysis of attributes on synthesis. We show the synthesis samples from either conventional or polarimetric thermal images on both datasets. Given probe (thermal) images and estimated attributes from the gallery (visible) image, our proposed method can generate attribute preserving (visible) images.

TABLE X: The Tufts Face Database [40] verification performance comparisons among the baseline methods and the proposed method.

Method	AUC	EER
Raw	66.73%	38.13%
Pix2Pix [23]	69.73% ± 0.92%	35.83% ± 0.59%
Pix2PixBEGAN [23], [1]	68.89% ± 0.51%	36.88% ± 0.43%
CycleGAN [70] (supervised)	71.93% ± 1.94%	34.16% ± 1.70%
GAN-VFS ‡ [62]	73.78% ± 0.46%	32.32% ± 0.53%
CRN + CL ‡ [34], [6]	74.90% ± 0.56%	31.71% ± 0.54%
Multi-AP-GAN (ours)	<b>75.86% ± 0.88%</b>	<b>31.14% ± 0.74%</b>
Multi-AP-GAN (GT) (ours)	<b>77.38% ± 0.98%</b>	<b>29.94% ± 0.79%</b>

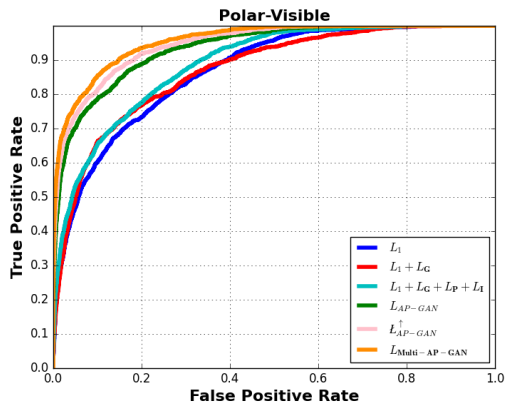


Fig. 17: The ROC curves corresponding to the ablation study.

translation mapping exactly guided by the visual attributes.

## VI. DISCUSSION

The proposed Multi-AP-GAN approach generates better quality visible images and as a result obtains improved cross-modal verification performance compared to previous GAN-based approaches. This can be attributed to the fact that Multi-AP-GAN uses a better generator which is guided by visual attributes. The multi-scale generator mitigates the

‡results are obtained after re-implementation due to the limited code availability.

\*features are extracted: [https://github.com/TreBlEn/InsightFace\\_Pytorch](https://github.com/TreBlEn/InsightFace_Pytorch)

receptive-field limitation of the convolutional operation by leveraging the features corresponding to images at multiple scales. In addition, visual attributes provide complementary semantic information for better synthesis. GAN-based methods such as GAN-VFS [62], Multi-stream GAN [63] and Pix2Pix [23] are single-scale generators and do not exploit such facial semantic information during synthesis.

Though our method performs reasonably well on three datasets, there are some limitations which we hope to overcome in our future work. Our model requires paired thermal and visible face images for training, which is laborious and expensive. Hence, an unsupervised synthesis method that does not require paired data is needed. Another limitation of our approach is that it does not work well on extreme pose variations. We are currently developing a new method that can deal with this pose issue in heterogeneous face recognition. We also plan to further investigate the impact of metabolic and physiologic variability in thermal facial signatures on synthesis and subsequent recognition performance.

## VII. CONCLUSION

We propose a novel Attribute Preserving Generative Adversarial Network (Multi-AP-GAN) structure for thermal-to-visible face verification via synthesizing photo-realistic visible face images from the corresponding thermal (polarimetric or conventional) images with extracted attributes. Rather than use only image-level information for synthesis and verification, we take a different approach in which semantic facial attribute information is also fused during training and testing. Quantitative and visual experiments evaluated on a real thermal-visible dataset demonstrate that the proposed method achieves state-of-the-art performance compared with other existing methods. In addition, an ablation study is developed to demonstrate the improvements obtained by different combination of loss functions.

## ACKNOWLEDGMENT

This work was supported by the Defense Forensics & Biometrics Agency (DFBA). The authors would like to thank Mr. Tom Cantwell and Ms. Michelle Giorgilli for their guidance

and extensive discussions on this work. The authors would like to express their appreciation to ODN/IARPA, as well as Chris Nardone, Marcia Patchan, and Stergios Papadakis at the JHU Applied Physics Laboratory for enabling ARL's participation in the 2018 IARPA Odin Program data collection, and Mathew Thielke for his help in collecting the data.

## REFERENCES

- [1] D. Berthelot, T. Schumm, and L. Metz. Began: boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.
- [2] T. Bourlai and L. A. Hornak. Face recognition outside the visible spectrum. *Image and Vision Computing*, 55:14 – 17, 2016.
- [3] T. Bourlai, N. Kalka, A. Ross, B. Cukic, and L. Hornak. Cross-spectral face verification in the short wave infrared (swir) band. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 1343–1347. IEEE, 2010.
- [4] P. Buddharaju, I. T. Pavlidis, P. Tsiamyrtzis, and M. Bazakos. Physiology-based face recognition in the thermal infrared spectrum. *IEEE transactions on pattern analysis and machine intelligence*, 29(4):613–626, 2007.
- [5] J. C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, March 2016.
- [6] N. Damer, F. Boutros, K. Mallat, F. Kirchbuchner, J.-L. Dugelay, and A. Kuijper. Cascaded generation of high-quality color visible face images from thermal captures. *arXiv preprint arXiv:1910.09524*, 2019.
- [7] T. de Freitas Pereira, A. Anjos, and S. Marcel. Heterogeneous face recognition using domain specific units. *IEEE Transactions on Information Forensics and Security*, 14(7):1803–1816, 2018.
- [8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.
- [9] X. Di and V. M. Patel. Face synthesis from visual attributes via sketch using conditional vaes and gans. *arXiv preprint arXiv:1801.00077*, 2017.
- [10] X. Di, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel. Polarimetric thermal to visible face verification via self-attention guided synthesis. In *2019 International Conference on Biometrics (ICB 2019)*, 2019.
- [11] X. Di, H. Zhang, and V. M. Patel. Polarimetric thermal to visible face verification via attribute preserved synthesis. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10. IEEE, 2018.
- [12] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He. Dual variational generation for low shot heterogeneous face recognition. In *Advances in Neural Information Processing Systems*, pages 2674–2683, 2019.
- [13] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2016.
- [14] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–326, 2016.
- [15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [16] K. P. Gurton, A. J. Yuffa, and G. W. Videen. Enhanced facial recognition for thermal imagery using polarimetric imaging. *Opt. Lett.*, 39(13):3857–3859, Jul 2014.
- [17] A. Gyaourova, G. Gebis, and I. Pavlidis. Fusion of infrared and visible images for face recognition. In *European Conference on Computer Vision*, pages 456–468. Springer, 2004.
- [18] R. He, J. Cao, L. Song, Z. Sun, and T. Tan. Adversarial cross-spectral face completion for nir-vis face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [19] R. He, X. Wu, Z. Sun, and T. Tan. Learning invariant deep representation for nir-vis face recognition. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [20] R. He, X. Wu, Z. Sun, and T. Tan. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1761–1773, 2018.
- [21] S. Hu, J. Choi, A. L. Chan, and W. R. Schwartz. Thermal-to-visible face recognition using partial least squares. *JOSA A*, 32(3):431–442, 2015.
- [22] S. Hu, N. J. Short, B. S. Riggan, C. Gordon, K. P. Gurton, M. Thielke, P. Gurram, and A. L. Chan. A polarimetric thermal database for face recognition research. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 119–126, 2016.
- [23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.
- [24] J. Choi, S. Hu, S. S. Young, L. S. Davis. Thermal to visible face recognition. In *Proc.SPIE*, pages 8371 – 8371 – 10, 2012.
- [25] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.
- [26] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [27] B. Klare and A. K. Jain. Heterogeneous face recognition: Matching nir to visible light images. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 1513–1516. IEEE, 2010.
- [28] B. F. Klare and A. K. Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1410–1422, 2013.
- [29] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570, 2015.
- [30] J. Lezama, Q. Qiu, and G. Sapiro. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6807–6816. IEEE, 2017.
- [31] A. Litvin, K. Nasrollahi, S. Escalera, C. Ozcinar, T. B. Moeslund, and G. Anbarjafari. A novel deep network architecture for reconstructing rgb facial images from thermal for face recognition. *Multimedia Tools and Applications*, 78(18):25259–25271, 2019.
- [32] Y. Lu, Y.-W. Tai, and C.-K. Tang. Attribute-guided face generation using conditional cyclegan. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 282–297, 2018.
- [33] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 5188–5196, 2015.
- [34] K. Mallat, N. Damer, F. Boutros, A. Kuijper, and J.-L. Dugelay. Cross-spectrum thermal to visible face recognition based on cascaded image synthesis. In *ICB 2019, 12th IAPR International Conference On Biometrics, 4-7 June, Crete, Greece, Crete, GRÈCE, 06 2019*.
- [35] K. Mallat and J.-L. Dugelay. A benchmark database of visible and thermal paired face images across multiple variations. In *International Conference of the Biometrics Special Interest Group, BIOSIG 2018, Darmstadt, Germany, September*, LNI, pages 199 – 206. GI / IEEE.
- [36] R. Mechrez, I. Talmi, and L. Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783, 2018.
- [37] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.
- [38] F. Nicolo and N. A. Schmid. Long range cross-spectral face recognition: matching swir against visible light images. *IEEE Transactions on Information Forensics and Security*, 7(6):1717–1726, 2012.
- [39] O. Nikisins, K. Nasrollahi, M. Greitans, and T. B. Moeslund. Rgb-d-t based face recognition. In *2014 22nd International Conference on Pattern Recognition*, pages 1716–1721, 2014.
- [40] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani, and X. Yuan. A comprehensive database for benchmarking imaging systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3):509–520, 2020.
- [41] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.
- [42] N. Pham and R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–247. ACM, 2013.
- [43] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2017.
- [44] R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J. C. Chen, V. M. Patel, C. D. Castillo, and R. Chellappa. Deep learning for understanding faces: Machines may be just as good, or better, than humans. *IEEE Signal Processing Magazine*, 35(1):66–83, Jan 2018.

- [45] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16*, pages 1060–1069. JMLR.org, 2016.
- [46] B. S. Riggan, N. J. Short, and S. Hu. Thermal to visible synthesis of face images using multiple regions. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.
- [47] B. S. Riggan, N. J. Short, S. Hu, and H. Kwon. Estimation of visible spectrum faces from polarimetric thermal faces. In *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*, pages 1–7. IEEE, 2016.
- [48] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [49] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [50] N. Short, S. Hu, P. Gurram, and K. Gurton. Exploiting polarization-state information for cross-spectrum face recognition. In *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*, pages 1–6. IEEE, 2015.
- [51] N. Short, S. Hu, P. Gurram, K. Gurton, and A. Chan. Improving cross-modal face recognition using polarimetric imaging. *Optics letters*, 40(6):882–885, 2015.
- [52] N. Short, S. Hu, P. Gurram, K. Gurton, and A. Chan. Improving cross-modal face recognition using polarimetric imaging. *Opt. Lett.*, 40(6):882–885, Mar 2015.
- [53] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.
- [54] T. Bourlai, A. Ross, C. Chen, L. Hornak. A study on using mid-wave infrared images for face recognition. volume 8371, pages 8371 – 8371 – 13, 2012.
- [55] L. Wang, V. Sindagi, and V. Patel. High-quality facial photo-sketch synthesis using multi-adversarial networks. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 83–90. IEEE, 2018.
- [56] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.
- [57] A. Wesley, P. Buddharaju, R. Pienta, and I. Pavlidis. A comparative analysis of thermal and visual modalities for automated facial expression recognition. In *International Symposium on Visual Computing*, pages 51–60. Springer, 2012.
- [58] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.
- [59] X. Wu, H. Huang, V. M. Patel, R. He, and Z. Sun. Disentangled variational representation for heterogeneous face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9005–9012, 2019.
- [60] S. Xie and Z. Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.
- [61] A. Yu, H. Wu, H. Huang, Z. Lei, and R. He. Lamp-hq: A large-scale multi-pose high-quality database for nir-vis face recognition. *arXiv preprint arXiv:1912.07809*, 2019.
- [62] H. Zhang, V. M. Patel, B. S. Riggan, and S. Hu. Generative adversarial network-based synthesis of visible faces from polarimetric thermal faces. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 100–107, Oct 2017.
- [63] H. Zhang, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel. Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. *International Journal of Computer Vision: Special Issue on Deep Learning for Face Analysis*, 2019.
- [64] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1947–1962, 2019.
- [65] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.
- [66] T. Zhang, A. Wiliem, S. Yang, and B. Lovell. Tv-gan: Generative adversarial network based thermal to visible face recognition. In *2018 International Conference on Biometrics (ICB)*, pages 174–181, 2018.
- [67] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang. Mdnnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436, 2017.
- [68] Z. Zhang, Y. Xie, and L. Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6199–6208, 2018.
- [69] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.
- [70] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.