Synthesis of High-Quality Visible Faces from Polarimetric Thermal Faces using Generative Adversarial Networks

He Zhang · **Benjamin S. Riggan** · **Shuowen Hu** · **Nathaniel J. Short** · **Vishal M. Patel**

Abstract The large domain discrepancy between faces captured in polarimetric (or conventional) thermal and visible domain makes cross-domain face verification a highly challenging problem for human examiners as well as computer vision algorithms. Previous approaches utilize either a two-step procedure (visible feature estimation and visible image reconstruction) or an input-level fusion technique, where different Stokes images are concatenated and used as a multi-channel input to synthesize the visible image given the corresponding polarimetric signatures. Although these methods have yielded improvements, we argue that input-level fusion alone may not be sufficient to realize the full potential of the available Stokes images. We propose a Generative Adversarial Networks (GAN) based multi-stream feature-level fusion technique to synthesize high-quality visible images from prolarimetric thermal images. The proposed network consists of a generator sub-network, constructed using an encoder-decoder network based on dense residual blocks, and a multi-scale discriminator sub-network. The generator network is trained by optimizing an adversarial loss in addition to a perceptual loss and an identity preserving loss to enable photo realistic generation of visible images while preserving discriminative characteristics. An extended dataset consisting of polarimetric thermal facial signatures of 111 subjects is also introduced. Multiple experiments evaluated on different experimental protocols demonstrate that the proposed method achieves state-of-the-art performance. Code will be made available at https://github.com/hezhangsprinter.

He Zhang
Department of ECE
Rutgers, The State University of New Jersey
94 Brett Road, Piscataway, NJ 08854
E-mail: he.zhang92@rutgers.edu

Benjamin S. Riggan
U.S. Army Research Laboratory
E-mail: benjamin.s.riggan.civ@mail.mil

Shuowen Hu
U.S. Army Research Laboratory
E-mail: shuowen.hu.civ@mail.mil

Nathaniel J. Short
Booz Allen Hamilton
E-mail: short_nathaniel@bah.com

Vishal M. Patel
Department of ECE, Johns Hopkins University
E-mail: vpatel36@jhu.edu

1 Introduction

Face is one of the most widely used biometrics for person recognition. Various face recognition systems have been developed over the last two decades. Recent advances in machine learning and computer vision methods have provided robust frameworks that achieve significant gains in performance of face recognition systems [54], [50], [2]. Deep learning methods, enabled by the vast improvements in processing hardware coupled with the ubiquity of face data, have led to significant improvements in face recognition accuracy, particularly in unconstrained face imagery [42], [3], [43].

Even though these methods are able to address many challenges and have even achieved human-expert level performance on challenging databases such as the low-resolution, pose variation and illumination variation to some extent [55], [39], [2], [6], [42], they are specifically designed for recognizing face images that are collected in the visible spectrum. Hence, they often do not perform well on the face images

captured from other domains such as thermal [46], [71], [14], [15], infrared [24], [34] or millimeter wave [9], [10] due to significant phenomenological differences as well as a lack of sufficient training data.
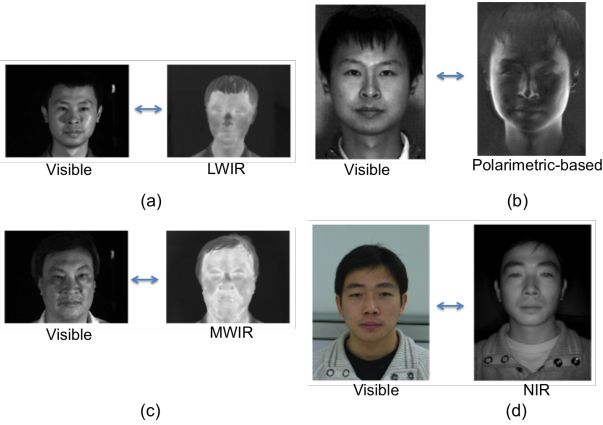

Fig. 1: Examples of (a) visible-LWIR pair [44], (b) visible-polarimetric pair [51], (c) visible-MWIR pair [44], and (d) visible-NIR pair [44].

Thermal imaging has been proposed for night-time and low-light face recognition when external illumination is not practical due to various collection considerations. The infrared spectrum can be divided into a reflection dominated region consisting of the near infrared (NIR) and shortwave infrared (SWIR) bands, and an emission dominated thermal region consisting of the midwave infrared (MWIR) and longwave infrared (LWIR) bands [45]. In particular, recent works have been proposed to use the polarization-state information of thermal emissions to enhance the performance of thermal face recognition [15], [46], [51], [71]. It has been shown that polarimetric-thermal images capture geometric and textural details of faces that are not present in the conventional thermal facial imagery [51]. As a result, the use of polarization-state information can improves cross-spectrum recognition performance over using intensity-only information from conventional thermal imagers.

Thermal face imagery, which can be acquired passively at night, but are currently not maintained in biometric-enabled watch lists, must be compared with visible-light face images for interoperability with existing biometric face databases. Distributional/domain differences between thermal and visible images makes thermal-to-visible face recognition very challenging (see Figure 1). Various methods have been developed in the literature to bridge this gap for cross-domain (i.e., heterogeneous) face recognition [48, 14, 49, 45, 18]. In particular, methods that synthesize visible faces from thermal facial signatures have gained traction in recent years [46], [71]. One of the advantages of face synthesis is that once the face images are synthesized in the visible domain, any off-the-shelf face matching algorithm can be used to

match the synthesized image to the gallery of visible images.

A polarimetric signature/image is defined here as consisting of three Stokes images as its three channels, analogous to the RGB channels in visible color imagery. Previous approaches utilize either a two-step procedure (visible feature estimation and visible image reconstruction) [46] or a fusion technique where different Stokes images are concatenated and used as a multi-channel input [71] to synthesize the visible image. Though these methods are able to effectively synthesize photo-realistic visible face images, the results are still far from optimal. One possible reason lies in that these methods concatenate the Stokes images into a single input sample without any additional attempts to capture multi-channel information inherently present in the different Stokes (modalities) images from the thermal infrared band. Hence, in order to efficiently leverage the multi-modal information provided by the polarimetric thermal images, we propose a novel multi-stream feature-level fusion method for synthesizing visible images from thermal domain using recently proposed Generative Adversarial Networks [11].

The proposed GAN-based network consists of a generator, a discriminator sub-network and a deep guided sub-network (see Figure 2). The generator is composed of a multi-stream encoder-decoder network based on dense-residual blocks, the discriminator is designed to capture features at multiple-scales for discrimination and the deep guided sub-net aims to guarantee that the encoded features contain geometric and texture information to recover the visible face. To further enhance the network's performance, it is guided by perceptual loss and an identity preserving loss in addition to adversarial loss. Once the face images are synthesized, any off-the-shelf face recognition and verification networks trained on the visible-only face data can be used for matching. Figure 3 illustrates the differences between visible and polarimetric thermal images. In addition, this figure also presents the photo-realistic and identity-preserving results obtained from our proposed method.

In addition to developing a novel face synthesis network, we also collected an extended dataset containing of visible and polarimetric facial signatures from 111 subjects. A subset of this dataset consisting data from 60 subjects was described in [15]. This extended polarimetric thermal facial dataset is available to computer vision and biometrics researchers to facilitate the development of cross-spectrum and multi-spectrum face recognition algorithms.

To summarize, this paper makes the following contributions.

1. A novel face synthesis framework based on GAN is proposed which consists of a multi-stream generator and multi-scale discriminator.
2. To embed the identity information into the objective function and make sure that the synthesized face images are
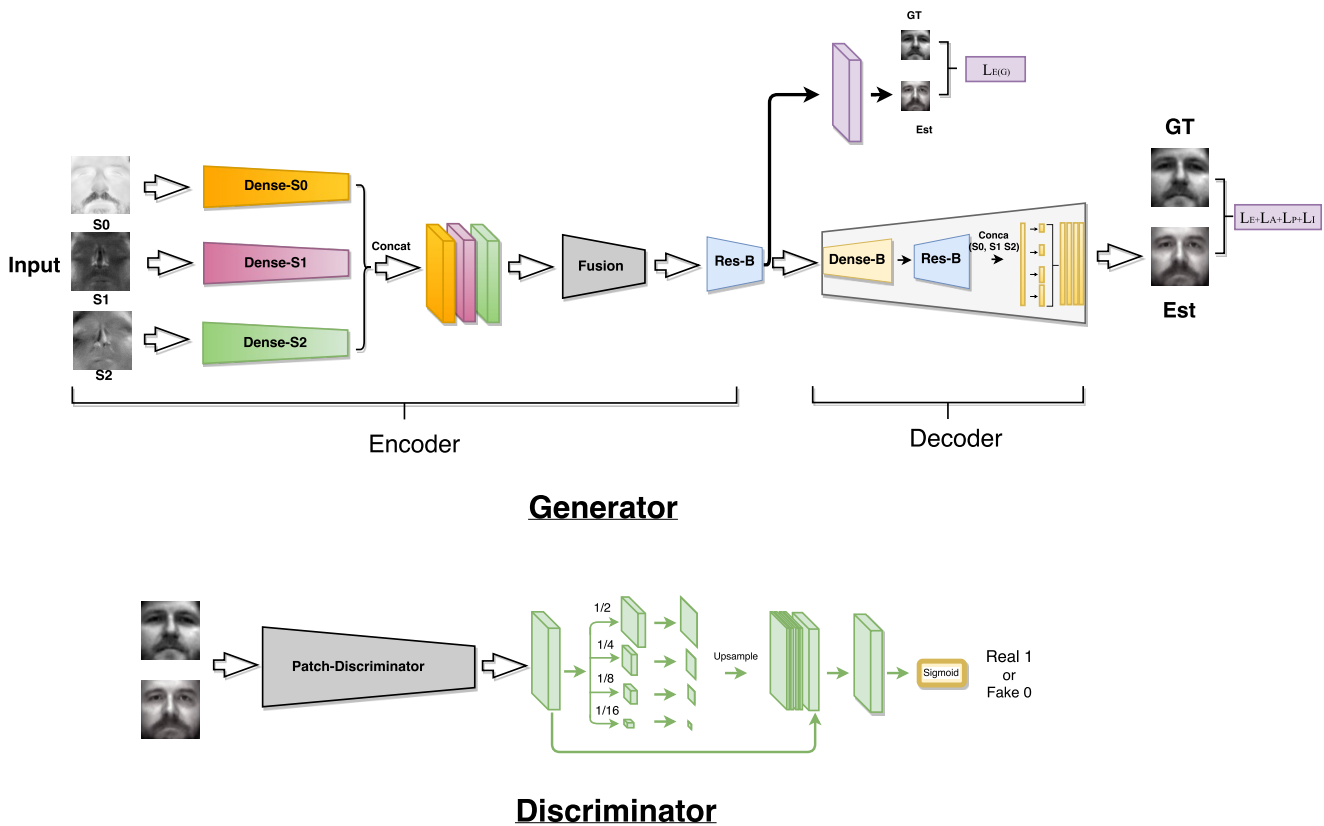
Fig. 2: An overview of the proposed GAN-based multi-stream encoder-decoder network. The generator contains a multi-stream feature-level fusion encoder-decoder network. In addition, a deep-guided subnet is stacked at the end of the encoding part. The discriminator is composed of a multi-scale patch-discriminator structure.

photo-realistic, a refined loss function is proposed for training the network.

3. An extended dataset consisting of visible and polarimetric data from 111 subjects is collected.
4. Detailed experiments are conducted to demonstrate improvements in the synthesis results. Further, three ablation studies are conducted to verify the effectiveness of iterative synthesis and various loss function.

Compared to the previous approaches for polarimetric thermal to visible face synthesis, this is a completely new work and one of the first approaches that proposes to use GANs for synthesizing high-quality visible faces from polarimetric thermal faces. Our previous work [71] also addresses the same problem but there are several notable differences: 1) The newly proposed method includes a novel multi-stream densely-connected network to transfer the polarimetric thermal facial signatures into the visible domain. 2) Feature-leavel fusion of different Stokes images is explored in this work to demonstrate the effectiveness of leveraging multiple polarimetric modalities for visible face synthesis. 3) We introduce an extended dataset that includes 111 subjects. 4) A novel multi-scale discriminator is introduced

that leverages the information from different scales to decide whether the given image is real or fake.

The paper is organized as follows. In Section 2, we review related works and give a brief background on GANs and polarimetric thermal imaging. Details of the proposed multi-stream feature-level fusion method are discussed in Section 3. In Section 4, we describe the extended polarimetric thermal face dataset. Experiments and results are presented in Section 5 and Section 6 concludes the paper with a brief summary and discussion.

2 Background and Related Work

In this section, we give a brief overview of polarimetric thermal imaging and GANs. Then, we review some related works on heterogeneous face recognition algorithms.

2.1 Polarimetric Thermal Imaging

Polarimetric thermal imaging uses advanced materials and sensor technology to measure the polarization state of light.
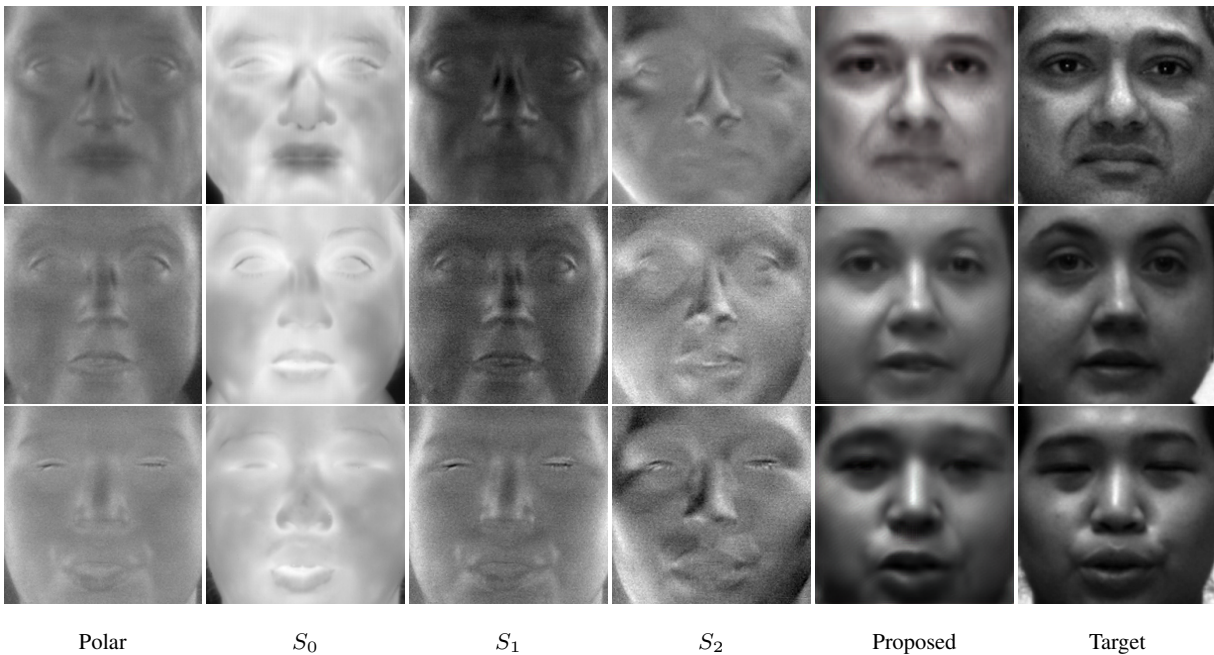
| Polar | S_0 | S_1 | S_2 | Proposed | Target |

Fig. 3: Sample results of the proposed method. (a) Input Polarimetric image. (b) Input S_0 image. (c) Input S_1 image. (d) Input S_2 image (e) Results from the proposed method, and (f) Target image.

While traditional imaging exploits the intensity of light, polarimetric imaging exploits the orientation of its oscillations. Natural visible light exhibits no preferred polarization state. If natural light is either transmitted across a boundary from one medium to another, or is reflected by the boundary (i.e., the material is opaque), a preferential polarization state (usually linear) may occur.

This induced polarization change is a directional quantity and is a function of the angle between the surface normal and the transmitted/reflected ray. For example, unpolarized sunlight reflecting off an air-water interface results in an induced linear polarization state that is orthogonal to the plane of reflection, as defined by the surface normal and the reflected ray. A similar phenomena exists when considering light energy in the "thermal" infrared (IR) part of the spectrum, e.g., MidIR (3-5μm) and/or LWIR (8-12μm). For induced polarization in the thermal IR, the radiation is treated as either emitted and/or reflected from a surface boundary. It is this interaction at the boundary that results in an induced net linear polarization state, similar to situation seen for visible light. By capturing this thermal radiance using an IR polarimetric camera, one can exploit the additional polarization based information and reconstruct a 3D surface from a 2D polarimetric image.

Polarimetric imaging sensors capture polarization-state information through optical filtering of light at different polarizations. This is traditionally done using a rotating element [56] (i.e., division of time), but other approaches exist, such as micro-grid polarizers [56] (i.e., division of focal plane array). In essence, polarization-state information is captured at four orientations, I_0, I_{90}, I_{45}, and I_{135}. The I_0 and I_{90} measurements represent horizontal and vertical polarized light and I_{45} and I_{135} capture diagonally polarized light. A stack of 2-D images captured using a polarimeter is represented by Stokes images, as defined in [12], which contain geometric and textural features, such as edges of the nose and mouth as well as wrinkles. These Stokes images are illustrated in Figure 3 for three subjects with corresponding visible-spectrum facial signatures. The S_0 image is a total intensity polarimetric image and is representative of what a conventional thermal imager (i.e., without linear polarizer) would capture. S_1, and S_2 illustrate the additional details provided by polarimetric imaging. In this paper, we refer to *Polar* as the three channel polarimetric image with S_0, S_1 and S_2 as the three channels.

2.2 Generative Adversarial Networks

Generative Adversarial Networks were first proposed by Goodfellow et al. in [11] to synthesize realistic images by effectively learning the distribution of the training images. The authors adopted a game theoretic min-max optimization framework to simultaneously train two models: a generative model, G, and a discriminative model, D. Initially, the success of GANs was limited as they were known to be unstable to train, often resulting in artifacts in the synthesized images. Radford *et al.* in [41] proposed Deep Convolutional GANs (DCGANs) to address the issue of instability by including a set of constraints on their network design. Another limiting issue in GANs is that, there is no control on the modes

of data being synthesized by the generator in case of these unconditioned generative models. Mirza *et al.* [32] incorporated additional conditional information in the model, which resulted in effective learning of the generator. The use of conditioning variables for augmenting side information not only increased the stability in learning but also improved the descriptive power of the generator G [22]. Recently, researchers have explored various aspects of GANs such as training improvements [47] and use of task specific cost function [5]. Also, an alternative viewpoint for the discriminator function is explored by Zhao *et al.* [75] where they deviate from the traditional probabilistic interpretation of the discriminator model. The objective function of a conditional GAN is defined as follows

$$
\begin{aligned}
L_{cGAN}(G, D) = &E_{x,y \sim Pdata(x,y)}[\log D(x, y)] + \\
&E_{x \sim Pdata(x), z \sim p_z(z)}[\log(1 - D(x, G(x, z)))],
\end{aligned}
\tag{1}
$$

where y, the output image, and x, the observed image, are sampled from distribution $Pdata(x, y)$ and they are distinguished by the discriminator, D. While for the generated fake $G(x, z)$ sampled from distributions $x \sim Pdata(x), z \sim p_z(z)$ would like to fool D.

The success of GANs in synthesizing realistic images has led researchers to explore the GAN framework for numerous applications such as data augmentation [38], zero-shot learning [76], image inpainting [66], image dehazing [62, 70, 64], text-to-image translation [61], image-to-image translation [19, 73, 40], texture synthesis [20], crowd-counting [52] and generating outdoor scenes from attributes [22]. Isola *et al.* [19] proposed a general method for image-to-image translation using conditional adversar'ial networks. Apart from learning a mapping function, they argue that the network also learns a loss function, eliminating the need for specifying or designing a task specific loss function. Karacan *et al.* in [22] proposed a deep GAN conditioned on semantic layout and scene attributes to synthesize realistic outdoor scene images under different conditions. Recently, Jetchev *et al.* [20] proposed spatial GANs for texture synthesis. Different from traditional GANs, their input noise distribution constitutes a whole spatial tensor instead of a vector, thus enabling them to create architectures more suitable for texture synthesis.

2.3 Heterogeneous Face Recognition

Recently, there has been a growing number of approaches that bridge existing modality gaps in order to perform heterogeneous face recognition. These approaches focused on various scenarios of heterogeneous face recognition such as infrared-to-visible [27, 13, 53], thermal-to-visible [48, 45, 71, 18, 24], and sketch-to-visible [8, 36] [57]. Fundamentally, each approach seeks to either find a common latent subspace

in which corresponding faces from each domain are "close" in terms of some distance and non-corresponding faces are "far." or synthesize photo-realistic visible face given its corresponding input modality.

Klare and Jain [25] proposed an approach using kernel prototype similarities, where after geometric normalization and image filtering (e.g., Difference of Gaussian, Center-Surround Divisive Normalization [31], and Gaussian) and local features extraction (e.g., multi-scale local binary patterns, or MLBP, and scale invariant feature transform, or SIFT), the intra-domain kernel similarities are computed between source (or target) domain images and all training images from the source (or target) domain. These intra-domain kernel similarity, which are computed using the cosine kernel, provide relational vectors for source and target domain imagery to be compared, where the main idea is that the kernel similarity between two source domain images should be similar to the kernel similarity between two corresponding target domain images.

Yi *et al.* [65] leverage the use of multi-modal Restricted Boltzmann Machines (RBMs) [33] to learn a shared representation, and for NIR-to-visible face recognition. Here, they learn shared representation using the multi-modal RBMs locally for each patch. However, since heterogeneity is only addressed locally, they further reduce the modality gap by performing Hetra-component analysis (HCA) [28] for the holistic image representation. Hetero-component analysis is based on the theory that most of the appearance differences between imaging modalities are captured in the top eigenvectors. Therefore, a common representation is given by removing the effects from the top eigenvectors. This was shown to achieve excellent performance for NIR-to-visible face recognition. However, it is unclear how well this would work for an emissive infrared band, such LWIR, where facial signatures are very different than in visible or NIR bands due to phenomenology.

Riggan *et al.* [44] proposed coupled auto-associative network for learning common representation between thermal and visible face images. The authors optimize two sparse auto-encoders jointly, such that (1) information within each modality is preserved and (2) inter-domain representations are similar for corresponding images. Although this approach demonstrated some success and robustness, the constraint to preserve information for the source domain is not a necessary condition as long as discriminability is maintained when learning common representation.

Hu *et al.* [14] applied a one-versus-all framework using partial least squares classifiers on Histogram of Oriented Gradients (HOG) features. For each classifier, they introduce the concept of adding cross-domain negative samples (i.e., thermal samples from a different subject) for added robustness. Later, Riggan *et al.* [45] proposed the using of a coupled neural network and a discriminative classifier for

enhance conventional thermal-to-visible face recognition and polarimetric thermal-to-visible framework.

While the described methods individually introduce custom approaches for reducing the modality gaps between two imaging domains, there are two fundamental concerns: (1) how to adjudicate matches when returned to an analyst, and (2) how to leverage recent advances in visible spectrum face recognition. Therefore, Riggan *et al.* [46] proposed a way to synthesize a visible image from both conventional thermal and polarimetric thermal. This approach used a convolution neural network to extract features from a conventional or polarimetric thermal image and then mapped those features to a corresponding visible representation using a deep perceptual mapping [30], where this representation in inverted back to the imaging domain using the forward convolutional neural network model. One potential concern is the piece-wise nature of this synthesis method. Later, built on the success of generative adversarial networks [11], Zhang *et al.* [71] improved the synthesis results by proposing an end-to-end conditional generative adversarial network (CGAN) approach, which is optimized via a newly introduced identity-preserving loss, to synthesis a corresponding visible image given a thermal image. This approach demonstrated results that were photo-realistic and discriminative.

3 Proposed Method

As discussed earlier, a polarimetric sample consists of three different Stokes images (S_0, S_1 and S_2), where S_0 represents the conventional thermal image and S_1 and S_2 represent the horizontal/vertical and diagonal polarization-state information, respectively. Unlike traditional three-channel RGB images where each channel contains different spectral information, the S_0, S_1, S_2 images contain different geometric and texture information. For example, as shown in the first row of Figure 3, S_0 is able to capture the mustache information, which is not captured in S_1 and S_2. On the other hand, S_0 does not capture some of the other texture and geometric details such as wrinkles and the shape of the mouth, which are well-preserved in S_1 and S_2. In other words, the Stokes images individually capture different facial features and when combined together they provide complementary information. Hence, it is important to fully utilize the information from all three Stokes images to effectively synthesize a visible face image.

Previous methods have attempted to utilize this information by exploiting input level fusion, where three Stokes images are concatenated together as a three-channel input [46,71]. Even though the three-channel concatenation in the input level is able to generate better visible face results by bringing in the geometric and texture differences preserved in these three modalities as compared with using just a single Stokes image as input (eg. S_0), the results are still far

from optimal [71]. A potential reason is that input level fusion or mere concatenation of different Stokes images may not be sufficient enough to exploit the different geometric and texture information present in these modalities. [1]. To efficiently address this problems and generate better photo-realistic visible face images, a multi-stream feature-level fusion structure is proposed in this paper. Specifically, different encoder structures are leveraged to encode each Stokes image separately and then the embedded features from each encoder are fused together via a fusion block for further visible face reconstruction (i.e. decoding).

Synthesizing photo-realistic visible images from polarimetric images (or even any single Stokes image) is an extremely challenging problem due to information differences caused by phenomenology between polarimetric thermal images and visible images. As shown in Figure 3, polarimetric thermal images fail to capture fine details such as edges and gradients as compared to visible images. Due to the absence of these sharp details in the polarimetric images, synthesizing visible images from them requires joint modeling of the images from these two domains. To efficiently leverage the training samples and guarantee better convergence with less gradient vanishing for such joint modeling, a novel dense residual structure is proposed in this paper. Furthermore, a multi-scale patch-discriminator is utilized to classify between real and synthesized images at multiple scales. By performing the discrimination at multiple scales, we are able to effectively leverage contextual information in the input image, resulting in better high-frequency details in the reconstructed image.

To summarize, we propose a multi-stream feature-level fusion GAN structure (see Figure 2) which consists of the following components:
(1) Multi-stream densely-connected encoder.
(2) Deep guidance sub-network.
(3) Single-stream dense residual decoder.
(4) Multi-scale discriminator.
In what follows, we describe these components in detail.

3.1 Multi-stream Feature-level Fusion Generator

The proposed feature-level fusion method is inspired by the face dis-entangled representation work proposed by Peng *et al.* and Tran *et al.* in [39,55,37], where the encoded feature representations are explicitly disentangled into separate parts representing different facial priors such as identity, pose and gender. Rather than leveraging the supervised label information to enforce the disentangling factor in the embedded features, each encoder structure in the proposed

---

[1] Input level fusion can be regarded as an extreme case for low-level feature fusion, where low-level features (from shallow layers) often preserve edge information rather than semantic mid-level or high-level class-specific information [67].

method inherently learns to characterize different geometric and texture information that is captured in the Stokes images. This information is then combined with a residual block-based fusion network, followed by a decoder network, consisting of a dense network and a residual network, to reconstruct visible domain faces from the fused feature maps. Furthermore, a deep-guided sub-network is leveraged at the end of the encoding part to ensure that the encoded features preserve geometric and texture information.

Multi-stream Densely-connected Encoding. The encoder consists of three streams of sub-networks, with each sub-network having the same structure. [2]. Each stream processes a particular input Stokes image. Basically, each stream is composed of a convolutional layer with kernel size 4, stride 2 and zero-padding 1, rectified linear unit (ReLU) and a 2x2 max-pooling operator at the front followed by three level dense-blocks [16] [3]. Each layer $\mathbf{D}_j$ in a dense block can be represented as

$$\mathbf{D}_j = T(cat[D_1, D_2, ..., D_{j-1}]), \tag{2}$$

where $T(\cdot)$ indicates the combination of Batch Normalization (BN) [17], rectified linear unit (ReLU) and Convolution operator. Figure 4 gives an overview of a single stream in the multi-stream densely-connected encoding.
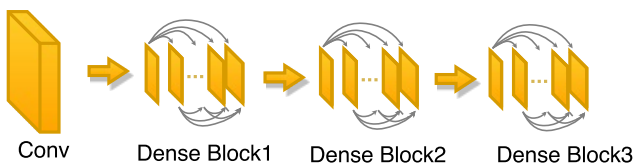

Fig. 4: Overview of a single stream in the multi-stream densely-connected encoding part.

There are three levels of densely connected blocks where each separate level contains 12, 16, or 24 dense-blocks. In addition, there exist one transition down layer which is composed of a 1×1 convolution layer followed by 2×2 average pooling between two contiguous dense blocks. All the parameters in each stream in the encoder can be initialized using the pre-trained Dense-net 121 structure [16]. Each dense-block contains a 3×3 convolution layer with ReLU and Batch normalization. In addition, the output of each dense block is concatenated with the input of the corresponding dense block. Once we calculated features from all three streams, we concatenate together from all three branches along the depth (channel) dimension. Feature maps from each of the three streams are of size $C \times H \times W$.

[2] Weights are not shared among each stream.
[3] Feature map size (width and height) in each level is same.

These feature maps are concatenated and are forwarded to the residual-fusion block, which consists of a res-block with 1×1 convolution layer. Then, the output of the residual-fusion block is regarded as the input for two different branches. To guarantee that the learned features contain geometric and textural facial information, a deep guidance sub-network [60] is introduced at the end of the encoding part as one branch. The deep guided sub-network is part of the network that is branching out from the end of the encoder. This sub-network is composed of a 1×1 convolution layer followed by the non-linear function, Tanh. Hence, the output of the guided sub-network will be a three-channel RGB image with size 16×16 if the input size is 256×256. In addition, the decoder is regarded as another branch discussed below.

Dense-Residual Decoder. The fused feature representations are then fed into a decoder network that is based on dense-residual decoding blocks. Specifically, the decoder contains five dense-residual blocks, where each dense-residual block contains a dense block, a transition up block and two residual blocks. Each dense block has the same structure as the dense block described in the encoder. Each transition up layer is composed of a 1×1 convolution layer followed by a bilinear up-sampling layer. Each residual blocks contains two 3×3 convolution layer connected by the ReLU function and Batch normalization. Once the feature maps are up-sampled to the original resolution (input resolution, e.g. 256×256), these learned features are concatenated with the three input Stokes images. Finally, a multi-level pyramid pooling block is adopted at the end of the decoding part to make sure that features from different scales are embedded in the final result. This is inspired by the use of global context information in classification and segmentation tasks [74, 69]. Rather than taking very large pooling size to capture more global context information between different objects [74], more 'local' information is leveraged here. Hence, a four-level pooling operation with sizes 1/32, 1/16, 1/8 and 1/4 are used. Then, features from all four levels are up-sampled to the original feature size and are concatenated back with the original feature maps before the final estimation. Figure 5 gives an overview of the proposed dense-residual decoder.

Specifically, the final multi-scale pyramid pooling structure contains a four scale down-sampling operator followed by a 1x1 convolution layer with one-channel output functioning as depth-wise dimension reduction. Then, all four scale one-channel feature maps are concatenated with the corresponding input of the multi-scale pooling structure by up-sampling to the input-feature resolution. Finally, the concatenated features are fed into a 3×3 convolution layer followed by a Tanh layer.
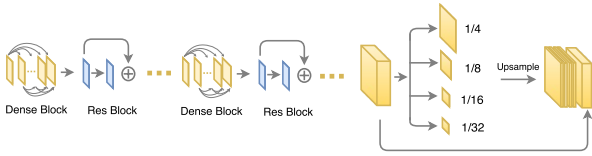
Fig. 5: Overview of the dense-residual decoding part.

3.2 Multi-scale Discriminator

To ensure the synthesized visible faces are indistinguishable from real images while preserving high-frequency details, a learned multi-scale patch-discriminator sub-network is designed to decide if each input image (to the discriminator) is real or fake. Similar to the structure that was proposed in [19], a convolution layer with batch normalization and Leaky ReLU [29] activation are used as the basis throughout the patch-discriminator part. Basically, the patch-discriminator consists of the following structure:

$$CBL(K_2)\text{-}CBL(2K_2)\text{-}CBL(4K_2)\text{-}CBL(8K_2)$$

where, $CBL(K_2)$ is a set of K_2-channel convolution layers followed by batch normalization and Leaky ReLU [29]. Then, a multi-scale pooling module, which pools features at different scales, is stacked at the end of the discriminator. The pooled features are then upsampled and concatenated, followed by a 1×1 convolution and a sigmoid function to produce a probability score normalized between 0 and 1. The proposed discriminator sub-network, D, is shown at the bottom of Figure 2.

3.3 Loss Functions

It is well-known that the use of Euclidean loss, L_E, alone often results in blurry results. Hence, to overcome this and to discriminate the generated visible face images from their corresponding ground truth, an adversarial loss function is employed. Even though the use of adversarial loss can generate more reasonable results compared to the L_E loss, as shown in [71], these results contain undesirable facial artifacts. To address this issue and generate visually pleasing results, perceptual loss is incorporated in our work. The perceptual loss is computed using a pre-trained VGG-16 models as discussed in [21, 68, 72, 26].

Since the ultimate goal of the our proposed synthesis method is to guarantee that human examiners or face verification systems can identify the person given his/her synthesized face images, it is also important to involve the discriminative information into consideration. Similar to the perceptual loss, we propose an identity-preserving loss that is evaluated on a certain layer of the fine-tuned VGG-Polar model. The VGG-Polar model is fine-tuned using the visi-

ble images with their corresponding labels from the newly introduced Polarimetric Visible database.

The proposed method contains the following loss functions: the Euclidean L_2 loss enforced on the reconstructed visible image, the $L_{E(G)}$ loss enforced on the guidance part, the adversarial loss to guarantee more sharp and indistinguishable results, the perceptual loss to preserve more photo realistic details and the identity loss to preserve more discriminative information for the outputs. The overall loss function is defined as follows

$$L_{\text{all}} = L_2 + L_{2(G)} + \lambda_A L_A + \lambda_P L_P + \lambda_I L_I, \qquad (3)$$

where L_2 denotes the Euclidean loss, $L_{2(G)}$ denotes the Euclidean loss on the guidance sub-network, L_A represents the adversarial loss, L_P indicates the perceptual loss and L_I is the identity loss. Here, λ_A, λ_P and λ_I are the corresponding weights.

The L_2 and the adversarial losses are defined as follows:

$$L_2, L_{2(G)} = \sum_{w,h} \|\phi_G(S_0, S_1, S_2)^{w,h} - Y_t^{w,h}\|_2, \qquad (4)$$

$$L_A = -\log(\phi_D(\phi_G(S_0, S_1, S_2)), \qquad (5)$$

where S_0, S_1 and S_2 are the three different input Stokes images, Y_t is the ground truth visible image, $W \times H$ is the dimension of the input image, ϕ_G is the multi-stream feature-fusion generator sub-network G and ϕ_D is the multi-scale discriminator sub-network D.

As the perceptual loss and the identity losses are evaluated on a certain layer of the given CNN model, both can be defined as follows:

$$L_{P,I} = \sum_{c_i, w_i, h_i} \|V(\phi_G(S_0, S_1, S_2))^{c_i, w_i, h_i} - V(Y_t)^{c_i, w_i, h_i}\|_2,$$

$$(6)$$

where Y_t is the ground truth visible image, ϕ_E is the proposed generator, V represents a non-linear CNN transformation and C_i, W_i, H_i are the dimensions of a certain high level layer V, which differs for perceptual and identity losses.

4 Polarimetric Thermal Face Dataset

A polarimetric thermal face database of 111 subjects is used for this study, which expanded on the previously released database of 60 subjects (described in detail in Hu *et al.*, 2016 [15]). The database used in this study therefore consisted of the 60-subject database collected at the U.S. Army Research Laboratory (ARL) in 2014-2015 (referred to as

Volume 1 hereinafter), and a 51-subject database collected at a Department of Homeland Security test facility (referred to as Volume 2 hereinafter). While the participants of the Volume 1 collect consisted exclusively of ARL employees, the participants of the Volume 2 collect were recruited from the local community in Maryland, resulting in more demographic diversity. Note that this extended databased is available upon request.

4.1 Sensors

The sensors employed to collect Volume 1 and Volume 2 were the same, consisting of a polarimetric LWIR imager and visible cameras. The LWIR polarimetric was developed by Polaris Sensor Technologies, and is based on a division-of-time spinning achromatic retarder (SAR) design which incorporated a spinning phase-retarder in conjunction with a linear wire-grid polarizer. This system has a spectral response range of 7.5-11 μm, and employed a Stirling cooler with a mercury telluride focal plane array (640×480 pixel array format). Data was recorded at 60 frames per second, using a lens with a field of view (FOV) of $10.6° \times 7.9°$. Four Basler Scout GigE cameras with different lens (ranging from $5°$ to $53°$) were used for Volume 1, consisting of two grayscale cameras (model # scA640-70gm; 659×494 pixel FPA) and two color cameras (model # scA640-70gc; 658×492 pixel FPA) to generate visible facial imagery at different resolutions. For Volume 2, a single Basler Scout color camera with a zoom lens was used, adjusted to produce the same facial resolution as the polarimeter.
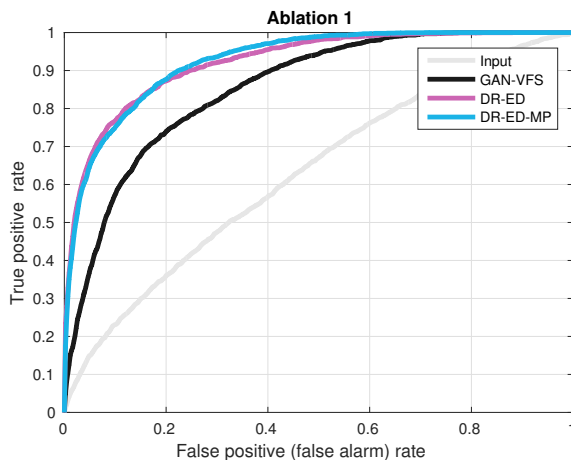
4.2 Dataset

The dataset protocols for Volume 1 and Volume 2 were approved by the respective Institutional Review Boards (IRBs) where each collection occurred. The Volume 1 collection involved two experimental conditions: range and expressions. Acquisitions were made at distances of 2.5 m, 5 m, and 7.5 m. At each range, a 10 second video sequence was first collected of the subject with a neural expression, and then a 10 second "expressions" sequence was collected as the subject counted out loud numerically from one upwards, which induced a continuous range of motions of the mouth and, to a lesser extent the eyes. In the experimental setup for Volume 1, a floor lamp was placed 1 m in front of the subject at each range to provide additional illumination.

Table 1: The average PSNR (dB), SSIM, EER and AUC results corresponding to different methods for **Ablation 1**.

	I-Polar	GAN-VFS [71]	DR-ED	DR-ED-MP
PSNR (dB)	11.74	18.07	18.28	**18.80**
SSIM	0.4625	0.7047	0.7128	**0.7194**
EER	41.51%	22.45%	16.51%	**15.67%**
AUC	62.93%	86.10%	91.67%	**92.55%**

The data collection setup used for Volume 2 matched that of Volume 1. However, no floor lamp was employed in the Volume 2 collect, as the DHS test facility had sufficient illumination. Furthermore, Volume 2 data was collected at a single range of 2.5 m, due to time limitations since the polarimetric face acquisition was part of a broader collection.

4.3 Preprocessing

The raw polarimetric thermal imagery underwent several preprocessing steps. First, a two-point non-uniformity correction (NUC) was applied on the raw data using software provided by Polaris Sensor Technologies and calibration data collected with a Mikron blackbody prior to each session. Images were sampled/extracted from the polarimetric thermal sequences. Bad pixels in the extracted images were identified, and those pixel intensities corrected via a median filter. To crop and align the facial imagery, three fiducial points (centers of the eyes, base of the nose) were first manually annotated, and an affine transform was used to normalize each face to canonical coordinates. Facial imagery was finally cropped to $m \times n$ pixels, and saved as 16-bit PNG files. The visible imagery required neither non-uniformity correction nor bad pixel correction. The same steps were used to crop and align the visible images, which were then saved as 16-bit grayscale PNG files.
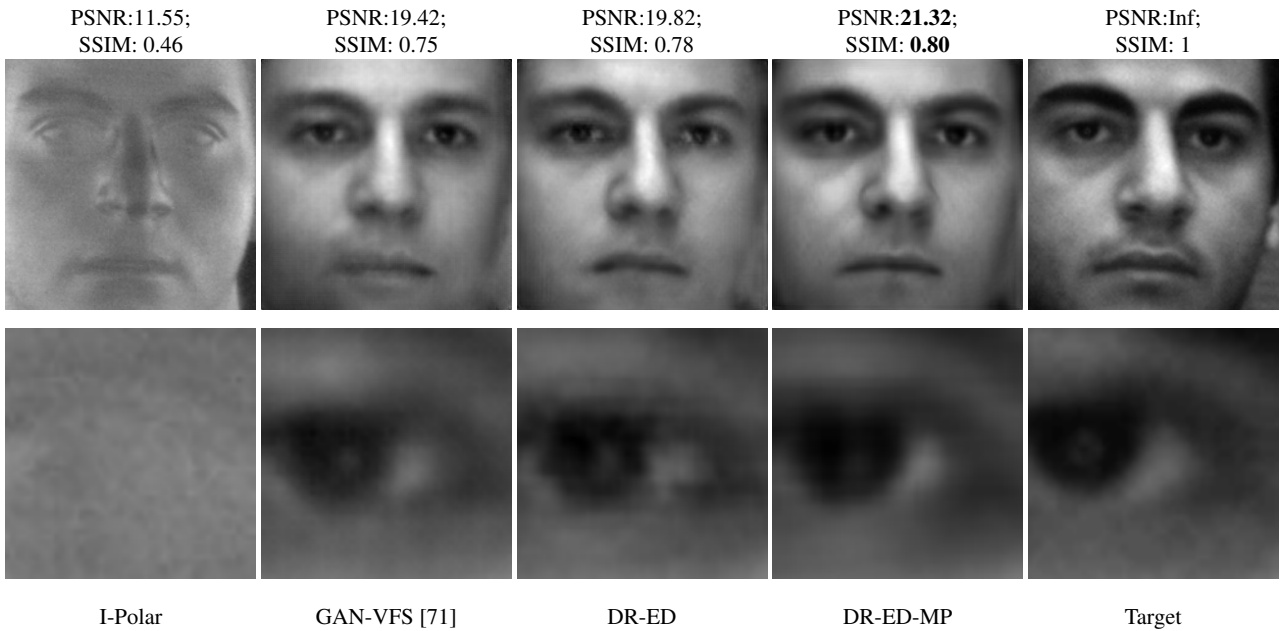

Fig. 6: The ROC curves corresponding to **Ablation 1**.

PSNR:11.55; SSIM: 0.46	PSNR:19.42; SSIM: 0.75	PSNR:19.82; SSIM: 0.78	PSNR:**21.32**; SSIM: **0.80**	PSNR:Inf; SSIM: 1
I-Polar	GAN-VFS [71]	DR-ED	DR-ED-MP	Target

Fig. 7: Sample results of for the **Ablation 1**. It can be observed that the dense-resisual encoder-decoder structure is able to generate better visible results and the introduced multi-level pooling module is able to preserve better structure information. Detail discussions can be found in Sec 5.2.

4.4 Experimental Protocols

Even though there exist several conventional thermal-visible pair databases [7,4], they lack the availability of the corresponding polarization state information such as S_1 and S_2. Hence, an extended database, which contains polarimetric (S_0, S_1, S_2) and visible image pairs from 111 subjects is used for evaluation in this paper. Following the protocol defined in [46,71], sample pairs corresponding to range 1 (baseline and expression) are used for comparisons. In particular, two different protocols are defined in this paper for further research. To be consistent with previous methods [46,71], the first protocol is defined as follows:

Protocol 1: The protocol 1 is evaluated on Volume 1, which contains 60 subjects, 30 subjects from Volume 1 with eight samples for each subject (in total 240 sample pairs) are used as training samples, denoted as *Train1*. Similarly, the remaining 30 subjects with eight samples for each subject (in total 240 sample pairs) are used as testing samples, denoted as *Protocol1*. All the training and testing samples are randomly chosen from the overall 60 subjects. Results are evaluated on five random splits. In Protocol 1, each split contains around 28800 pairs of templates on average (1080 positive and 27720 negative).

Protocol 2: Different from Protocol 1, the newly introduced and extended dataset with 111 subjects is used for training and testing, where 85 subjects with eight samples for each subject are randomly chosen as training samples (in total 680 sample pairs), denoted as *Train2* and the other 26 sub-

jects are used as testing (in total 208 sample pairs), denoted as *Protocol2*. As before, results are evaluated on five random splits. In Protocol 2, each split on average contains around 21632 pairs of templates (936 positive and 20696 negative).

These protocols and splits will be made publicly available to the research community.

5 Experimental Results

In this section, we demonstrate the effectiveness of the proposed approach by conducting various experiments on the two defined protocols for the new polarimetric thermal dataset as described above. Once the visible images are synthesized using the proposed method, deep features can be extracted from these images using any one of many pre-trained CNNs such as VGG-face [35], Light-CNN [59], or GoogleNet [63]. In this paper, we extract the features from the second last fully connected layer of the VGG-face network [35]. Finally, the cosine distance is used to calculate the scores. Results are compared with four state-of-the-art methods: Ben *et al.* [46], GAN-VFS [71], Pix2pix [19] and Pix2pix with BEGAN [19,1]. In addition, three ablation studies are conducted to demonstrate the effectiveness of different modules of the proposed method. Quality of the synthesized images is evaluated using Peak Signal-to-Noise Ratio (PSNR) and Structural SIMilarity (SSIM) index [58]. The face verification performance is evaluated using the receiver operating characteristic (ROC) curve, Area Under the Curve (AUC) and Equal Error Rate (EER) measures.
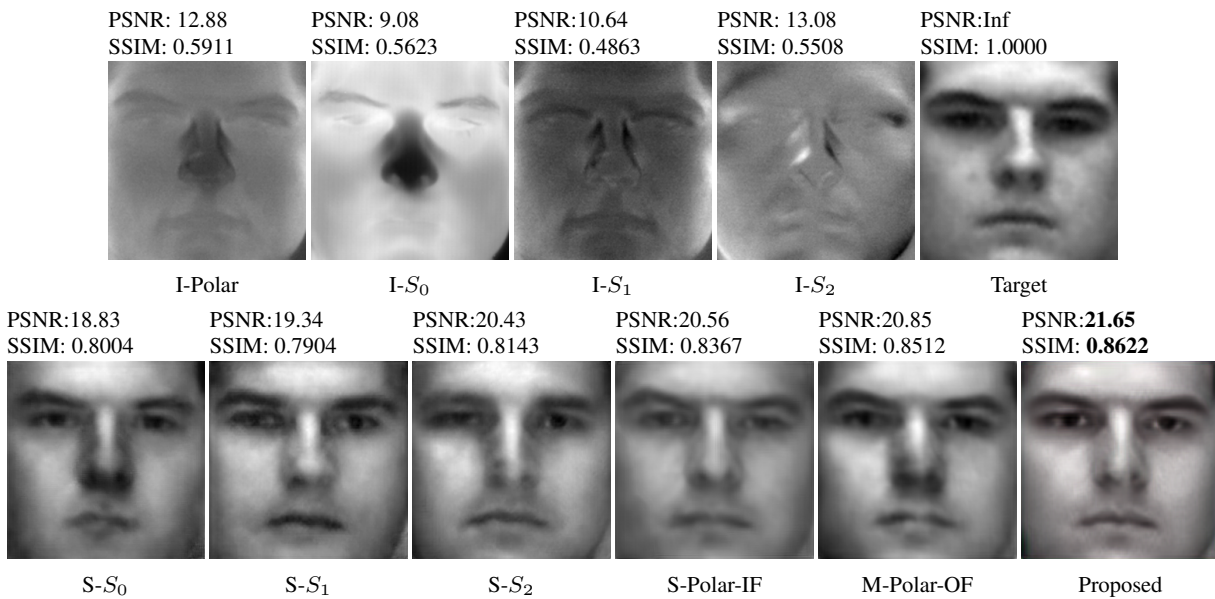
| PSNR: 12.88 | PSNR: 9.08 | PSNR:10.64 | PSNR: 13.08 | PSNR:Inf |
| SSIM: 0.5911 | SSIM: 0.5623 | SSIM: 0.4863 | SSIM: 0.5508 | SSIM: 1.0000 |

| I-Polar | I-S_0 | I-S_1 | I-S_2 | Target |

| PSNR:18.83 | PSNR:19.34 | PSNR:20.43 | PSNR:20.56 | PSNR:20.85 | PSNR:**21.65** |
| SSIM: 0.8004 | SSIM: 0.7904 | SSIM: 0.8143 | SSIM: 0.8367 | SSIM: 0.8512 | SSIM: **0.8622** |

| S-S_0 | S-S_1 | S-S_2 | S-Polar-IF | M-Polar-OF | Proposed |

Fig. 8: Sample results for **Ablation 2**. It can be observed that the proposed multi-stream feature-level fusion GAN is able to generate better results compared to input-level (S-Polar-IF), output-level fusion (M-Polar-OF) and also simply levering single Stokes modality. Detailed discussions can be found in Sec 5.2.

Table 2: The average PSNR (dB), SSIM, EER and AUC results corresponding to different methods for **Ablation 2**.

	I-Polar	S-S_0	S-S_1	S-S_2	S-Polar-IF	M-Polar-OF	Proposed
PSNR (dB)	11.74	17.34	17.03	17.17	18.80	18.87	**19.55**
SSIM	0.4625	0.6905	0.6852	0.6794	0.7194	0.7225	**0.7433**
EER	41.51%	23.18%	21.61%	21.56%	15.67%	15.90%	**11.78%**
AUC	62.93%	85.74%	86.64%	87.30%	92.55%	92.69%	**96.03%**

5.1 Implementation

The entire network is trained on a Nvidia Titan-X GPU. We choose $\lambda_A = 0.005$ for the adversarial loss, $\lambda_P = 0.8$ for the perceptual loss and $\lambda_I = 0.1$ for the identity loss. During training, we use ADAM [23] as the optimization algorithm with learning rate of 8×10^{-4} and batch size of 1 image. All the pre-processed training samples are resized to 256×256. The perceptual loss is evaluated on relu 1-1 and relu 2-1 layers in the pre-trained VGG [35] model. The identity loss is evaluated on the relu2-2 layer of the fine-tuned VGG-Polar model.

5.2 Ablation Study

In order to better demonstrate the effectiveness of the proposed feature-level fusion, the improvements obtained by different modules and the importance of different loss functions in the proposed network, three ablation studies are presented in this section. All the experiments in the first two ablation studies are optimized with the same loss function discussed in Eq (2).

Ablation 1

In the first ablation study, we demonstrate the effectiveness of different modules (eg. densely connected encoder-decoder structure) in our method by conducting the following experiments. All the experimental results are evaluated using **Protocol 1** based on the polrimetric images as input:

(a) **GAN-VFS**: The GAN network proposed in [71] with polarimetric images as inputs.
(b) **DR-ED**: A single stream dense-resisual encoder-decoder structure. [4]
(c) **DR-ED-MP**: A single stream dense-resisual encoder-decoder structure with multi-level pooling.

Table 3: The average PSNR (dB), SSIM, EER and AUC results corresponding to different methods for **Ablation 3**.

	I-Polar	L2	L2-GAN	L2-GAN-P	Our
PSNR (dB)	11.74	17.57	17.33	18.99	**19.55**
SSIM	0.4625	0.7088	0.7115	0.7352	**0.7433**
EER	41.51%	18.07%	13.23%	11.79%	**11.78%**
AUC	62.93%	90.89%	93.64%	95.64%	**96.03%**

One synthesis example corresponding to **Ablation 1** is shown in Figure 7. It can be observed from this figure (comparing second column with third column) that the overall performance improves after leveraging the newly introduced dense-residual encoder-decoder (DR-ED) structure. This can be clearly observed from the left part of the reconstructed

---

[4] Basically, this network is composed of one stream of the encoder part followed by the same decoder without multi-level pooling.
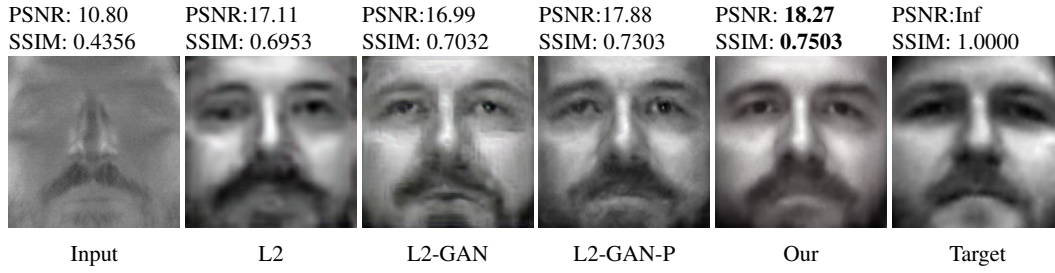
| PSNR: 10.80 | PSNR:17.11 | PSNR:16.99 | PSNR:17.88 | PSNR: **18.27** | PSNR:Inf |
| SSIM: 0.4356 | SSIM: 0.6953 | SSIM: 0.7032 | SSIM: 0.7303 | SSIM: **0.7503** | SSIM: 1.0000 |

| Input | L2 | L2-GAN | L2-GAN-P | Our | Target |

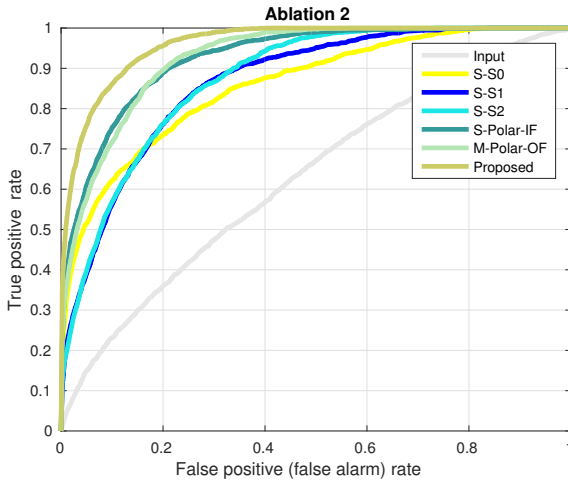**Fig. 9:** Sample results on different loss functions for **Ablation 3**.

Fig. 10: The ROC curves corresponding to **Ablation 2**.
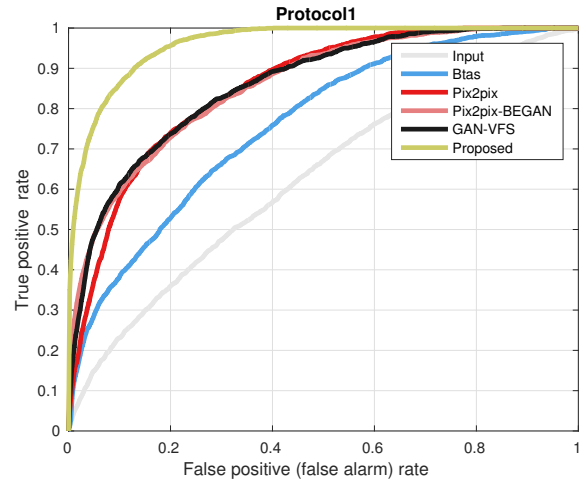
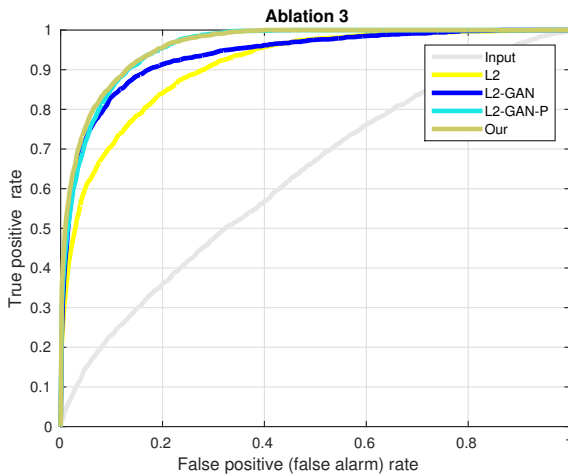**Fig. 12:** The ROC curves corresponding to *Protocol1*.

Fig. 11: The ROC curves corresponding to **Ablation 3**.

mouth. This essentially demonstrates the effectiveness of the proposed dense-residual encoder-decoder structure. Though the DR-ED is able to reconstruct better visible face, from the close-up of the left eye shown in the second row in Figure7 we observe that some structure information is missing. The multi-level pooling module at the end of the encoder-decoder structure overcomes this issue and preserves the the

overall eye structure. Quantitative results evaluated based on PSNR and SSIM [58], as shown in Table 1, also show similar results.

In addition to comparing the performance of the synthesized images in terms of SSIM and PSNR, we also compare the contribution of each module in face verification by plotting the ROC curves. The verification results are evaluated based on the cosine similarity using the deep features extracted from the pre-defined VGG-face model [35]. The results are shown in Figure 6. From the ROC curves, it can be clearly observed that the proposed dense-residual network with multi-level pooling can also provide some discriminative information. Similar results can also be observed from the EER and AUC comparisons, tabulated in Table 1.

Table 4: The PSNR, SSIM and EER and AUC results corresponding to *Protocol1*.

	I-Polar	Btas-2016 [46]	Pix2pix [19]	Pix2pix-BEGAN [19,1]	GAN-VFS [71]	Proposed
PSNR (dB)	11.74	16.12	16.79	17.55	18.07	**19.55**
SSIM	0.4625	0.6785	0.6490	0.7033	0.7041	**0.7433**
EER	41.51%	26.72%	22.61%	22.56%	23.19%	**11.78**%
AUC	62.93%	81.90%	85.14%	85.30%	85.89%	**96.03**%

| PSNR: 10.80 | PSNR:15.77 | PSNR:16.65 | PSNR:17.05 | PSNR:19.03 | PSNR:**19.52** | PSNR:Inf |
| SSIM: 0.4356 | SSIM: 0.6423 | SSIM: 0.6681 | SSIM: 0.6603 | SSIM: 0.7203 | SSIM: **0.7380** | SSIM: 1.0000 |

| PSNR: 10.23 | PSNR: 16.04 | PSNR:17.04 | PSNR:16.84 | PSNR:17.34 | PSNR:**18.05** | PSNR:Inf |
| SSIM: 0.4152 | SSIM: 0.6122 | SSIM: 0.6604 | SSIM: 0.6449 | SSIM: 0.6722 | SSIM: **0.7134** | SSIM: 1.0000 |

| I-Polar | Btas-2016 [46] | Pix2pix [19] | Pix2pix-BEGAN [19,1] | GAN-VFS [71] | Proposed | Target |
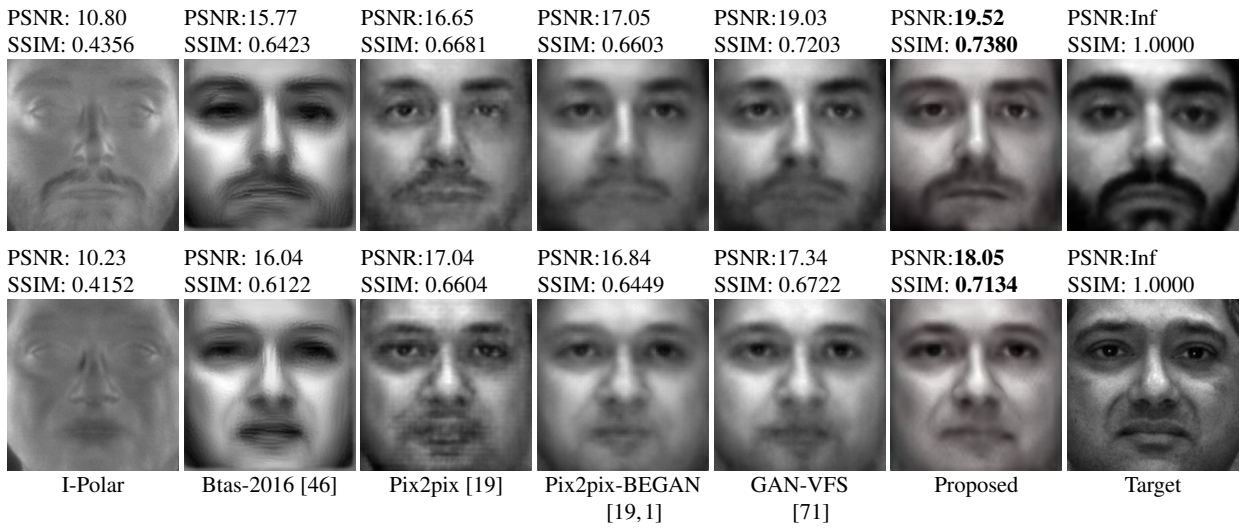
Fig. 13: Sample results compared with state-of-the-art methods evaluated on *Protocol1*.
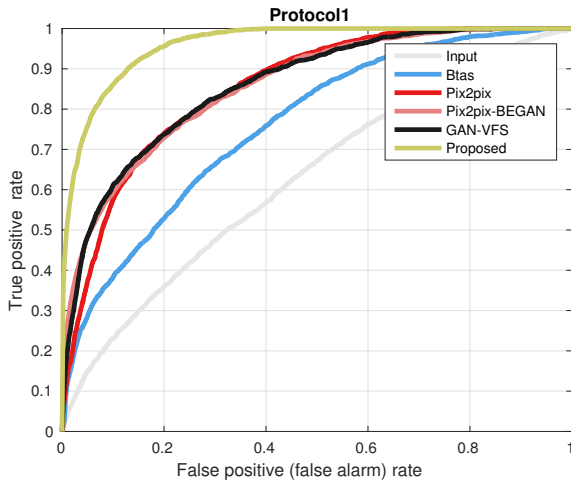
Fig. 14: The ROC curves corresponding to *Protocol1*.

Ablation 2

The second ablation study is conducted to demonstrate the effectiveness of the proposed feature level multi-model fusion by conducting experiments with the following baselines:

(a) **S-S_0:** Single stream dense-resisual encoder-decoder with the proposed structure with S_0 as the input.

(b) **S-S_1:** Single stream dense-resisual encoder-decoder with the proposed structure with S_1 as the input

(c) **S-S_2:** Single stream dense-resisual encoder-decoder with the proposed structure with S_2 as the input.

(d) **S-Polar-IF:** Single stream dense-resisual encoder-decoder with the proposed structure with Polar as the input (i.e. input level fusion). The S-Polar-IF model shares the exact same structure as DR-ED-ML as discussed in **Ablation 1**.

(e) **M-Polar-OF:** Multi stream dense-resisual encoder-decoder structure with output level fusion. The M-Polar-OF is basically composed of three stream dense-resisual encoder-decoder structure, where each stream shares the same structure with S-Polar-IF but with different input (S_0, S_1 and S_2) for each stream. Then, the output features from each stream are fused (concatenated) at the end of the decoding part to generate visible face images.

(f) **M-Polar-F-L2:** Multi-stream dense-resisual encoder-decoder with the proposed structure based on feature-level fusion optimized with L2 loss only.

(g) **M-Polar-F-L2-GAN:** Multi-stream dense-resisual encoder-decoder with the proposed structure based on feature-level fusion optimized with L2 and GAN loss.

(h) **M-Polar-F-L2-GAN-Perp:** Multi-stream dense-resisual encoder-decoder with the proposed structure based on feature-level fusion optimized with L2, GAN loss and perceptual loss.

(i) **Our (M-Polar-FF):** Multi-stream dense-resisual encoder-decoder with the proposed structure based on feature-level fusion with all the losses.

Sample results corresponding to **Ablation 2** is shown in Figure 8. It can be observed that just leveraging any one of the Stokes images as input is unable to fully capture the geometric and texture details of the whole face. For example, as shown in the first column second row in Figure 8, the nose is over-synthesized if just S_0 (representing conventional thermal imagery) is used. Leveraging input level fusion (just concatenating three modalities as three-channel input) S-Polar-IF enables better visible face with less undesired artifacts as compared to S-S_0, S-S_1 and S-S_2. Furthermore, the proposed multi-stream feature-level fusion structure is able to preserve more geometric facial details and is able to generate photo-realistic visible face images. Vi-
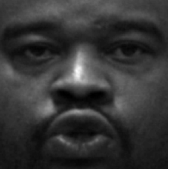
| PSNR: 10.36 | PSNR: 14.67 | PSNR:15.74 | PSNR:20.04 | PSNR:19.86 | PSNR: **21.66** | PSNR:Inf |
| SSIM: 0.4331 | SSIM: 0.6301 | SSIM: 0.6305 | SSIM: 0.7004 | SSIM: 0.7249 | SSIM: **0.7728** | SSIM: 1.0000 |

| PSNR: 10.25 | PSNR: 14.67 | PSNR:13.98 | PSNR:20.04 | PSNR:19.67 | PSNR: **23.55** | PSNR:Inf |
| SSIM: 0.4108 | SSIM: 0.6301 | SSIM: 0.5915 | SSIM: 0.7544 | SSIM: 0.7306 | SSIM: **0.8097** | SSIM: 1.0000 |

| PSNR: 11.01 | PSNR: 18.99 | PSNR:22.96 | PSNR:23.92 | PSNR:25.36 | PSNR:**26.32** | PSNR:Inf |
| SSIM: 0.4236 | SSIM: 0.7525 | SSIM: 0.8008 | SSIM: 0.8421 | SSIM: 0.8572 | SSIM: **0.8732** | SSIM: 1.0000 |

| PSNR: 9.13 | PSNR: 14.36 | PSNR:14.79 | PSNR:17.48 | PSNR:17.69 | PSNR:**22.45** | PSNR:Inf |
| SSIM: 0.3743 | SSIM: 0.6052 | SSIM: 0.6087 | SSIM: 0.7015 | SSIM: 0.7019 | SSIM: **0.8107** | SSIM: 1.0000 |

| I-Polar | Btas-2016 [46] | Pix2pix [19] | Pix2pix-BEGAN [19,1] | GAN-VFS [71] | Proposed | Target |

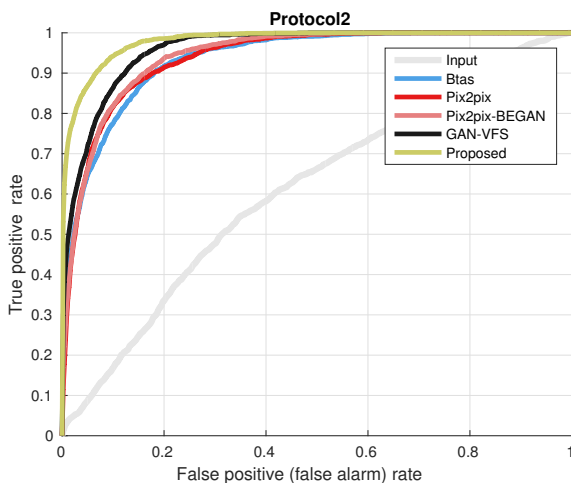**Fig. 15:** Sample results compared with state-of-the-art methods evaluated on *Protocol2*.

Fig. 16: The ROC curves corresponding to the *Protocol2*.

Quantitative results evaluated in terms of PSNR and SSIM are shown in Table 2. Results are also consistent with our visual comparison.

Similar to Ablation study 1, the face verification results are also used as a metric to evaluate the performnace of different fusion techniques. We plot the ROC curves corresponding to the different settings discussed above. The ROC curves are shown in Figure 10. Again, the verification results are evaluated based the cosine similarity using the deep features extracted from the VGG-face model [35] without fine-tuning. From the ROC curves, it can be clearly observed that the proposed multi-stream feature-level fusion can bring in more discriminative information as compared to input level or output level fusion.

Ablation 3

In the third ablation study, we demonstrate the effectiveness of different loss functions used in the proposed method (e.g. adversarial loss, perceptual loss and identity preserving loss) by conducting the following experiments. All the ex-

sual results also demonstrate the effectiveness of leveraging feature level fusion over input level or output level fusion.

Table 5: The PSNR, SSIM, EER and AUC results corresponding to *Protocol2*.

	I-Polar	Btas-2016 [46]	Pix2pix [19]	Pix2pix-BEGAN [19,1]	GAN-VFS [71]	Proposed
PSNR (dB)	10.88	15.82	17.82	18.28	18.58	**19.18**
SSIM	0.4467	0.6854	0.6828	0.7214	0.7283	**0.7340**
EER	40.87%	14.60%	13.49%	15.81%	11.42%	**7.99%**
AUC	61.27%	93.99%	93.46%	92.50%	95.96%	**98.00%**

perimental results are evaluated using **Protocol 1** based on the polarimetric images as the input:

(a) **L2**: The proposed architecture (M-Polar-FF) optimized with the L2 loss.

(b) **L2-GAN**: The proposed architecture optimized with the L2 loss and the adversarial loss.

(c) **L2-GAN-P**: The proposed architecture optimized with the L2 loss, the adversarial loss and the perceptual loss.

(d) **Our**: The proposed architecture optimized with the L2 loss, the adversarial loss, the perceptual loss and the identity-preserving loss.

Visual results corresponding to this ablation study are shown in Figure 12. It can be observed from the results that the L2 loss itself generates blurry faces and many details around the eyes and the mouth regions are missing. By involving the GAN structure in the proposed method, more details are being added to the results. But it can be observed that GAN itself produces images with artifacts. Introduction of the perceptual loss in the proposed framework is able to remove some of the artifacts and makes the results visually pleasing. Finally, the combination of all the losses is able to generate more reasonable results with better facial details.

To better demonstrate the effectiveness of different losses in the proposed method, we plot the ROC curves corresponding to the above four different network settings. The results are shown in Figure 11. All the verification results are evaluated on the deep features extracted from the VGG-face model [35] without fine-tuning. From the ROC curves, it can be clearly observed that even though the identity loss does not produce visually different results, it can bring in more discriminative information. The corresponding PSNR, SSIM values as well as the AUC and EER values are summarized in Table 3.

5.3 Comparison with State-of-the-Art Methods

To demonstrate the improvements achieved by the proposed method, it is compared against recent state-of-the-art methods [46, 19, 1, 71] on the new dataset. We compare quantitative and qualitative performance of different methods on the test images from the two distinct protocols *Protocol1* and *Protocol2* discussed earlier.

Sample results corresponding to Protocol 1 and Protocol 2 are shown in Figure 13 and Figure 15, respectively.

It can be observed from these figures, Pix2pix and Pix2pix-BEGAN introduce undesirable artifacts in the final reconstructed images.

The introduction of the perceptual loss in [71] is able to remove some of these artifacts and produce visually pleasing results. However, the synthesized images still lack some geometric and texture details as compared to the target image. In contrast, the proposed method is able to generate photo-realistic visible face images while better retaining the discriminative information such as the structure of mouth and eye. Quantitative results corresponding to different methods evaluated on both protocols are tabulated in Table 4 and Table5, showing that the proposed multi-stream feature-level fusion GAN structure is able to achieve superior performance.

Similar to the ablation study, we also propose to use the performance of face verification as a metric to evaluate the performance of different methods. Figure14 and Figure16 show the ROC curves corresponding to the two experimental protocols. The AUC and EER results are reported in Table 4 and Table 5. From these results, it can be clearly observed that the proposed method is able to achieve superior quantitative performance compared the previous approaches. These results highlight the significance of using a GAN-based approach to image synthesis.

6 Conclusion

We present a new multi-level dense-residual fusion GAN structure for synthesizing photo-realistic visible face images from the corresponding polarimetric data. In contrast to the previous methods that leverage input level fusion techniques to combine geometric and texture information from different Stokes image, we take a different approach where visual features extracted from different Stokes images are combined to synthesize the photo-realistic face images. Quantitative and qualitative experiments evaluated on a real polarimetric visible database demonstrate that the proposed method is able to achieve significantly better results as compared to the recent state-of-the-art methods. In addition, three ablation studies are performed to demonstrate the improvements obtained by the feature-level fusion methods, different modules and different loss functions in the proposed method. Furthermore, an extended polarimetric-visible database consisting of data from 111 subjects is also presented in this paper.

References

1. Berthelot, D., Schumm, T., Metz, L.: Began: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717 (2017)

2. Chen, J.C., Patel, V.M., Chellappa, R.: Unconstrained face verification using deep cnn features. In: Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on, pp. 1–9. IEEE (2016)

3. Chen, J.C., Ranjan, R., Sankaranarayanan, S., Kumar, A., Chen, C.H., Patel, V.M., Castillo, C.D., Chellappa, R.: Unconstrained still/video-based face verification with deep convolutional neural networks. International Journal of Computer Vision (2017). DOI 10.1007/s11263-017-1029-3

4. Chen, X., Flynn, P.J., Bowyer, K.W.: Ir and visible light face recognition. Computer Vision and Image Understanding **99**(3), 332–358 (2005)

5. Creswell, A., Bharath, A.A.: Task specific adversarial cost function. arXiv preprint arXiv:1609.08661 (2016)

6. Ding, H., Zhou, S.K., Chellappa, R.: Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In: Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, pp. 118–126. IEEE (2017)

7. Espinosa-Duró, V., Faundez-Zanuy, M., Mekyska, J.: A new face database simultaneously acquired in visible, near-infrared and thermal spectrums. Cognitive Computation **5**(1), 119–135 (2013)

8. Gao, F., Shi, S., Yu, J., Huang, Q.: Composition-aided sketch-realistic portrait generation. arXiv preprint arXiv:1712.00899 (2017)

9. Gonzalez-Sosa, E., Vera-Rodriguez, R., Fierrez, J., Patel, V.M.: Exploring body shape from mmw images for person recognition. IEEE Transactions on Information Forensics and Security **12**(9), 2078–2089 (2017)

10. Gonzalez-Sosa, E., Vera-Rodriguez, R., Fierrez, J., Patel, V.M.: Millimetre wave person recognition: hand-crafted vs. learned features. In: ISBA, pp. 1–7 (2017)

11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS, pp. 2672–2680 (2014)

12. Gurton, K.P., Yuffa, A.J., Videen, G.W.: Enhanced facial recognition for thermal imagery using polarimetric imaging. Optics letters **39**(13), 3857–3859 (2014)

13. He, R., Wu, X., Sun, Z., Tan, T.: Wasserstein cnn: Learning invariant features for nir-vis face recognition. arXiv preprint arXiv:1708.02412 (2017)

14. Hu, S., Choi, J., Chan, A.L., Schwartz, W.R.: Thermal-to-visible face recognition using partial least squares. JOSA A **32**(3), 431–442 (2015)

15. Hu, S., Short, N.J., Riggan, B.S., Gordon, C., Gurton, K.P., Thielke, M., Gurram, P., Chan, A.L.: A polarimetric thermal database for face recognition research. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 119–126 (2016)

16. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. arXiv preprint arXiv:1608.06993 (2016)

17. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of The 32nd International Conference on Machine Learning, pp. 448–456 (2015)

18. Iranmanesh, S.M., Dabouei, A., Kazemi, H., Nasrabadi, N.M.: Deep Cross Polarimetric Thermal-to-visible Face Recognition. ArXiv e-prints (2018)

19. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on (2016)

20. Jetchev, N., Bergmann, U., Vollgraf, R.: Texture synthesis with spatial generative adversarial networks. arXiv preprint arXiv:1611.08207 (2016)

21. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, pp. 694–711. Springer (2016)

22. Karacan, L., Akata, Z., Erdem, A., Erdem, E.: Learning to generate images of outdoor scenes from attributes and semantic layouts. arXiv preprint arXiv:1612.00215 (2016)

23. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

24. Klare, B., Jain, A.K.: Heterogeneous face recognition: Matching nir to visible light images. In: ICPR, pp. 1513–1516 (2010)

25. Klare, B.F., Jain, A.K.: Heterogeneous face recognition using kernel prototype similarities. IEEE transactions on pattern analysis and machine intelligence **35**(6), 1410–1422 (2013)

26. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2017)

27. Lezama, J., Qiu, Q., Sapiro, G.: 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), pp. 6807–6816. IEEE (2017)

28. Li, S., Yi, D., Lei, Z., Liao, S.: The casia nir-vis 2.0 face database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 348–353 (2013)

29. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models

30. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5188–5196 (2015)

31. Meyers, E., Wolf, L.: Using biologically inspired features for face processing. International Journal of Computer Vision **76**(1), 93–104 (2008)

32. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)

33. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp. 807–814 (2010)

34. Nicolo, F., Schmid, N.A.: Long range cross-spectral face recognition: Matching swir against visible light images. IEEE Transactions on Information Forensics and Security **7**(6), 1717–1726 (2012)

35. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proceedings of the British Machine Vision Conference (BMVC) (2015)

36. Peng, C., Gao, X., Wang, N., Tao, D., Li, X., Li, J.: Multiple representations-based face sketch–photo synthesis. IEEE transactions on neural networks and learning systems **27**(11), 2201–2215 (2016)

37. Peng, X., Feris, R.S., Wang, X., Metaxas, D.N.: A recurrent encoder-decoder network for sequential face alignment. In: European Conference on Computer Vision, pp. 38–56. Springer International Publishing (2016)

38. Peng, X., Tang, Z., Yang, F., Feris, R., Metaxas, D.: Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. arXiv preprint arXiv:1805.09707 (2018)

39. Peng, X., Yu, X., Sohn, K., Metaxas, D.N., Chandraker, M.: Reconstruction-based disentanglement for pose-invariant face recognition. In: Proceedings of the IEEE International Conference on Computer Vision (2017)

40. Perera, P., Abavisani, M., Patel, V.M.: In2i: Unsupervised multi-image-to-image translation using generative adversarial networks. arXiv preprint arXiv:1711.09334 (2017)

41. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)

42. Ranjan, R., Sankaranarayanan, S., Bansal, A., Bodla, N., Chen, J.C., Patel, V.M., Castillo, C.D., Chellappa, R.: Deep learning for understanding faces: Machines may be just as good, or better, than humans. IEEE Signal Processing Magazine **35**(1), 66–83 (2018). DOI 10.1109/MSP.2017.2764116

43. Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R.: An all-in-one convolutional neural network for face analysis. In: Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, pp. 17–24. IEEE (2017)

44. Riggan, B.S., Reale, C., Nasrabadi, N.M.: Coupled auto-associative neural networks for heterogeneous face recognition. IEEE Access **3**, 1620–1632 (2015). DOI 10.1109/ACCESS.2015.2479620

45. Riggan, B.S., Short, N.J., Hu, S.: Optimal feature learning and discriminative framework for polarimetric thermal to visible face recognition. In: Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on, pp. 1–7. IEEE (2016)

46. Riggan, B.S., Short, N.J., Hu, S., Kwon, H.: Estimation of visible spectrum faces from polarimetric thermal faces. In: Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on, pp. 1–7. IEEE (2016)

47. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: NIPS, pp. 2226–2234 (2016)

48. Sarfraz, M.S., Stiefelhagen, R.: Deep perceptual mapping for thermal to visible face recognition. arXiv preprint arXiv:1507.02879 (2015)

49. Sarfraz, M.S., Stiefelhagen, R.: Deep perceptual mapping for cross-modal face recognition. International Journal of Computer Vision **122**(3), 426–438 (2017)

50. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)

51. Short, N., Hu, S., Gurram, P., Gurton, K., Chan, A.: Improving cross-modal face recognition using polarimetric imaging. Opt. Lett. **40**(6), 882–885 (2015). DOI 10.1364/OL.40.000882

52. Sindagi, V.A., Patel, V.M.: Generating high-quality crowd density maps using contextual pyramid cnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1861–1870 (2017)

53. Song, L., Zhang, M., Wu, X., He, R.: Adversarial discriminative heterogeneous face recognition. AAAI (2018)

54. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in neural information processing systems, pp. 1988–1996 (2014)

55. Tran, L., Yin, X., Liu, X.: Disentangled representation learning gan for pose-invariant face recognition

56. Tyo, J.S., Goldstein, D.L., Chenault, D.B., Shaw, J.A.: Review of passive imaging polarimetry for remote sensing applications. Applied optics **45**(22), 5453–5469 (2006)

57. Wang, L., Sindagi, V.A., Patel, V.M.: High-quality facial photo-sketch synthesis using multi-adversarial networks. In: IEEE International Conference on Automatic Face and Gesture Recognition (2018)

58. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004)

59. Wu, X., He, R., Sun, Z., Tan, T.: A light cnn for deep face representation with noisy labels. IEEE Transactions on Information Forensics and Security **13**(11), 2884–2896 (2018)

60. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE international conference on computer vision, pp. 1395–1403 (2015)

61. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. arXiv preprint arXiv:1711.10485 (2017)

62. Xu, Z., Yang, X., Li, X., Sun, X., Harbin, P.: Strong baseline for single image dehazing with deep features and instance normalization

63. Yang, J., Ren, P., Zhang, D., Chen, D., Wen, F., Li, H., Hua, G.: Neural aggregation network for video face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4362–4371 (2017)

64. Yang, X., Xu, Z., Luo, J.: Towards perceptual image dehazing by physics-based disentanglement and adversarial training (2018)

65. Yi, D., Lei, Z., Li, S.Z.: Shared representation learning for heterogenous face recognition. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, vol. 1, pp. 1–7. IEEE (2015)

66. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. arXiv preprint arXiv:1801.07892 (2018)

67. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision, pp. 818–833. Springer (2014)

68. Zhang, H., Dana, K.: Multi-style generative network for real-time transfer. arXiv preprint arXiv:1703.06953 (2017)

69. Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

70. Zhang, H., Patel, V.M.: Densely connected pyramid dehazing network

71. Zhang, H., Patel, V.M., Riggan, B.S., Hu, S.: Generative adversarial network-based synthesis of visible faces from polarimetric thermal faces. International Joint Conference on Biometrics 2017 (2017)

72. Zhang, H., Sindagi, V., Patel, V.M.: Image de-raining using a conditional generative adversarial network. arXiv preprint arXiv:1701.05957 (2017)

73. Zhang, Z., Yang, L., Zheng, Y.: Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. arXiv preprint arXiv:1802.09655 (2018)

74. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1–8 (2017)

75. Zhao, J., Mathieu, M., LeCun, Y.: Energy-based generative adversarial network. arXiv preprint arXiv:1609.03126 (2016)

76. Zhu, Y., Elhoseiny, M., Liu, B., Elgammal, A.: Imagine it for me: Generative adversarial approach for zero-shot learning from noisy texts. arXiv preprint arXiv:1712.01381 (2017)