**Heterogeneous Face Frontalization via Domain Agnostic Learning**

Xing Di\(^1\) and Shuowen Hu\(^2\) and Vishal M Patel\(^1\)

\(^1\) Department of Electrical and Computer Engineering, Johns Hopkins University, Maryland, USA
\(^2\) U.S. Army DEVCOM Army Research Laboratory, Maryland, USA

**Abstract**—Recent advances in deep convolutional neural networks (DCNNs) have shown impressive performance improvements on thermal to visible face synthesis and matching problems. However, current DCNN-based synthesis models do not perform well on thermal faces with large pose variations. In order to deal with this problem, heterogeneous face frontalization methods are needed in which a model takes a thermal profile face image and generates a frontal visible face. This is an extremely difficult problem due to the large domain as well as large pose discrepancies between the two modalities. Despite its applications in biometrics and surveillance, this problem is relatively unexplored in the literature. We propose a domain agnostic learning-based generative adversarial network (DAL-GAN) which can synthesize frontal views in the visible domain from thermal faces with pose variations. DAL-GAN consists of a generator with an auxiliary classifier and two discriminators which capture both local and global texture discriminations for better synthesis. A contrastive constraint is enforced in the latent space of the generator with the help of a dual-path training strategy, which improves the feature vector’s discrimination. Finally, a multi-purpose loss function is utilized to guide the network in synthesizing identity-preserving cross-domain frontalization. Extensive experimental results demonstrate that DAL-GAN can generate better quality frontal views compared to the other baseline methods.

**I. INTRODUCTION**

Face recognition is a challenging problem which has been actively researched over the past few decades. In particular, deep convolutional neural network (DCNN) based methods have shown impressive performance improvements on various visible face recognition datasets. Numerous DCNN-based models have been proposed in the literature which can address key challenges that include expression, illumination and pose variations, image resolution, aging, etc. However, existing methods are specifically designed for recognizing face images that are collected in the visible spectrum. In applications such as night-time surveillance, we are faced with the scenario of identifying a face image acquired in the thermal domain by comparing it with a gallery of face images that are acquired in the visible domain. Existing DCNN-based visible face recognition methods will not perform well when directly applied to the problem of thermal to visible face recognition due to the significant distributional shift between the thermal and visible domains. In order to bridge this gap, various cross-domain face recognition algorithms have been proposed [9], [7], [6], [16], [15], [45], [46], [49], [31], [29], [5], [39], [38]. In particular, synthesis-based methods have gained a lot of traction in recent years [9]. Given a thermal face image, the idea is to synthesize the corresponding face image in the visible domain. Once the visible image is synthesized, any DCNN-based visible face recognition network can be leveraged for identification.

One of the main limitations of the existing synthesis-based models for cross-modal face recognition is that they do not perform well on thermal faces with large pose variations [9]. Face frontalization is an extensively studied problem in the computer vision and biometrics communities. Various methods have been developed for frontalization [14], [53], [40], [3], [41], [18], [19], [43], [44], [51], [52], [42], [50]. However, existing face frontalization methods’ performance degrades significantly on thermal faces since they are specifically designed for frontalizing visible face images. In order to deal with this problem, heterogeneous face frontalization methods are needed in which a model takes a thermal profile face image and generates a frontal visible face. This is an extremely difficult problem due to the large domain as well as large pose discrepancies between the two modalities. Despite its applications in biometrics and surveillance, this problem
is relatively unexplored in the literature.

We propose a domain agnostic learning-based generative adversarial network (DAL-GAN) [12] with dual-path training architecture which can synthesize frontal views in the visible domain from thermal faces with pose variations. Fig. 1 gives a simplified overview of the proposed heterogeneous frontalization framework. The model is trained on two flows: frontalization flow and feature-learning flow. The frontalization flow aims to synthesize a visible frontal face image from a thermal profile face image. The feature learning flow aims to learn domain-agnostic features that help in reconstructing better visible faces. DAL-GAN consists of a generator with an auxiliary classifier and two discriminators which capture both local and global texture discriminations for better synthesis. A contrastive constraint is enforced in the latent space of the generator with the help of a dual-path training strategy, which improves the feature vector’s discrimination. Finally, a multi-purpose loss function is utilized to guide the network in synthesizing identity-preserving cross-domain frontalization.

We conduct extensive experiments to demonstrate that DAL-GAN can generate better quality frontal views compared to the other baseline methods. Fig. 1 shows sample outputs from the proposed network.

In summary, this paper makes the following contributions.

- We propose a cross-spectrum face frontalization model via learning domain-agnostic features. To the best of our knowledge, this is one of the first models for cross-spectrum face frontalization.
- We introduce a dual-path network architecture for learning discriminative and domain-agnostic features. Features are obtained based on both a gradient reversal layer and the contrastive learning strategy.
- Extensive experiments and ablation studies are conducted to demonstrate the effectiveness of the proposed face frontalization method.

II. RELATED WORK

Face Frontalization: Recent face frontalization methods utilize either 2D/3D warping [14], [53], [22], [53], [52], stochastic modeling [40], [2], [1] or the generative models [3], [41], [18], [19], [43], [44], [51], [42], [50], [23]. For instance, Hassner et al.[14] proposed a single unmodified 3D surface model for frontalization. Sagonas et al.[40] proposed a model that jointly performs frontalization and landmark localization by solving a low-rank optimization. Rui et al.[19] developed TP-GAN with a two-path architecture for capturing both the global and local contextual information. Tran et al.[41] introduced the DR-GAN model which frontalsizes face images via disentangled representations. Hu et al.[18] introduced CAPG-GAN to synthesize frontal face images guided by the target landmarks. Zhao et al.[51] proposed the PIM model based on the domain adaptation strategy for pose invariant face recognition. Cao et al.[3] proposed the HF-PIM model which includes dense correspondence field estimation and facial texture map recovery. Similarly, Zhang et al.[50] developed the A3F-CNN model with the appearance flow constrain. Li et al.[28] proposed a frontalization model using a series of discriminators optimized by segmented face images. Yin et al.[44] proposed a self-attention-based generator is to integrate local features with their long-range dependencies for obtaining better frontalized faces. Wei et al.[42] proposed the FFWM model to overcome the illumination issue via flow-based feature warping.

Thermal-to-visible Synthesis: Various approaches have been proposed in the literature for thermal-to-visible face synthesis and matching [9], [16], [15], [45], [46], [49], [31], [29], [5], [39], [38], [8], [36], [20]. For instance, Hu et al.[16] developed a model based on partial least squares for heterogeneous face recognition. Riggan et al.[39] introduced the GAN-VFS model for cross-spectrum synthesis. Another TV-GAN model is introduced by Zhang et al.[49] which leverages the identity information via a classification loss. Riggan et al.[38] enhanced the image quality of the synthesized images by leveraging both global and local facial regions. Moreover, some models [47], [5] synthesize high-quality images by leveraging complementary information in multiple modalities. Recently, Mallat et al.[31] proposed a cascaded refinement network (CRN) model to synthesize visible images using a contextual loss. Litvin et al.[29] developed the FusionNet architecture to overcome the issue of model over-fitting on the RGB-T images [32]. Very recently, Di et al. introduced Multi-AP-GAN [9] and AP-GAN [10] models for improving the quality of the synthesized visible faces using facial attributes.

Note that the proposed model is different from the models discussed above. In our approach, a feature normalization-based generative adversarial network (GAN) is utilized to learn frontalization mapping. The generator is also optimized by learning domain-agnostic features through a gradient reversal layer (GRL) [11]. However, we find that simply learning domain-agnostic features is not enough for cross-spectrum frontalization, and learning discriminative features is also important. Therefore, semi-supervised contrastive learning is also implemented via a dual-path training strategy.

III. PROPOSED METHOD

Given a thermal face image x with a significant pose variation, our objective is to synthesize the corresponding frontal face image y in the visible domain. The generated frontal face image should be photo-realistic and identity-preserving. In order to address this cross-spectrum face frontalization problem, we propose a dual-path network architecture which learns domain-agnostic features via a contrastive learning strategy. In what follows, we present the proposed network in detail.

A. Networks Architecture

Fig. 2 gives an overview of the proposed dual-path architecture. Each path contains a generator F and two discriminators D0, D1. The two generators share weights.
Another domain classifier $C$ is utilized for learning domain-agnostic features. More details about the network architecture can be found in the supplementary file.

1) Generator: Each generator $F$ consists of an encoder-decoder structure where the encoder $E$ aims to extract domain-agnostic features from the input thermal face while the decoder $G$ aims to reconstruct a frontal face image in the visible domain. In this work, we adopt the generator architecture from DA-GAN [44] with the following modifications.

We remove all batch normalization layers in DA-GAN because they were originally introduced to eliminate the covariate shift. However, recent studies have shown that covariate shift often does not exist in GANs [25], [26], [27]. Instead we use feature vector equalization to prevent the escalation of feature magnitudes. Feature equalization also helps the network converge smoothly during training. Given the original feature vector $a_{m,n}$ at pixel location $(m, n)$, the equalization is defined as follows

$$b_{m,n} = a_{m,n} \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} a_{m,n}^i + \epsilon},$$

where $b_{m,n}$ is the normalized vector and $N$ is the total number of features. We set $\epsilon$ equal to $10^{-8}$. In our experiments, we find that feature equalization also allows us to use a higher initial learning rate which in turn helps in accelerating training.

2) Discriminator: A single discriminator may not be able to capture both global and local facial textures [19], [51]. Therefore, we employ two identical discriminators ($D_l, D_g$) to learn global and local discrimination respectively, as shown in Fig. 2. In particular, the local components of a frontal face $\hat{y}_f$ are extracted by a predefined off-the-shelf model $F_m$ [30]. In this work, the key components (eyes, nose, lips, brow) are extracted to learn the local facial structure and the entire face image is used for learning the global texture. Mathematically, we define a mask $M$ to extract the key components of a ground-truth visible image $y$ as:

$$M = F_m(y).$$

With the help of this mask, we can obtain local regions of a real/fake image by the element-wise product $\odot$ between the mask and the real/fake image. The local/global discriminators are optimized by the following loss function $L_{adv}^{l} / L_{adv}^{g}$ separately

$$L_{adv}^{g} = \mathbb{E}[D_g(y)] - \mathbb{E}[(D_g(\hat{y}^r))] - \lambda_{gp}\mathbb{E}[\|\nabla_y \cdot D_g(y^r)\|_2 - 1]^2,$$

$$L_{adv}^{l} = \mathbb{E}[D_l(M \odot y)] - \mathbb{E}[(D_l(M \odot \hat{y}))] - \lambda_{gp}\mathbb{E}[\|\nabla_M \odot y^r \cdot D_l(M \odot y^r)\|_2 - 1]^2,$$

$$L_{adv} = L_{adv}^{g} + \lambda_l \cdot L_{adv}^{l}.$$  

Here, $y^r$ is sampled uniformly along a straight line between a pair of real image $y_f$ and the generated image $\hat{y}_f$ [13]. The discriminators attempts to maximize this objective, while the generator attempts to minimize it. We set $\lambda_{gp} = 10, \lambda_l = 0.1$ in our experiments.

B. Objective Function

The loss function we use to train the proposed network consists of a combination of contrastive loss, gradient reversal layer-based domain adaptive loss and frontalization loss.

Frontalization Loss: To capture texture information between the real and fake images at multiple scales, we utilize a
multi-scale pixel loss as follows

$$L_{pixel} = \sum_{s=1}^{3} \| \hat{y}^s - y^s \|_1,$$  \hspace{1cm} (3)

where $s$ corresponds to the resolution scale. In our work, we utilize three different resolution scales: $32 \times 32$, $64 \times 64$ and $128 \times 128$.

Identity preserving synthesis is important for heterogeneous face recognition. To achieve this, we utilize the identity loss as follows

$$L_{id} = \| F_i(\hat{y}) - F_i(y) \|_2,$$  \hspace{1cm} (4)

where $F_i$ is the pre-trained VGGFace [34] model which is used to extract features. By minimizing the feature distance between the real and fake images, the generator is optimized to synthesize identity-preserving images.

In order to reduce the artifacts in the synthetic images, a total variation regularization $L_{tv}$ is also applied as that in [24]. Hence, the total loss for the frontalization flow is defined as follows

$$L_{front} = L_{pixel} + \lambda_1 \cdot L_{id} + \lambda_{adv} \cdot L_{adv} + \lambda_{tv} \cdot L_{tv},$$  \hspace{1cm} (5)

where $\lambda_1$, $\lambda_{adv}$, $\lambda_{tv}$ are parameters.

**Domain Adaptive Loss:** To learn domain-agnostic features, one domain classifier network $C$ with a gradient reversal layer [11] is employed in our network. The domain classifier network is a simple 3-layer multi-linear perceptron (MLP) neural network with the same equalization as in Eq (1). Given both thermal and visible face images, the classifier aims to correctly estimate which spectrum the input image belongs to. When back-propagating, the gradient reversal layer flips the gradients which leads to domain-agnostic features in the encoder. The network parameters $\theta_E$ corresponding to the encoder network are updated as follows

$$L_{cls} = \frac{1}{N} \sum_{i=0}^{N-1} -k_i \log(C(E(x_i))) - (1 - k_i) \log(1 - C(E(x_i)))$$

$$\theta_E = \theta_E - \mu \left( \frac{\partial L_{front}}{\partial \theta_E} - \lambda_{grl} \frac{\partial L_{cls}}{\partial \theta_E} \right),$$  \hspace{1cm} (6)

where $x_i$ is the $i$-th input sample, $\mu$ is the learning rate, $\theta_E$ and $\theta_E$ are the updated and original parameters, respectively.

$k_i$ is the domain index $[0, 1]$. $\lambda_{grl}$ is set equal to 0.01 in our experiments.

**Contrastive Loss:** Finally, the contrastive loss [4] is used to enhance the discriminability of the latent features. It is defined as follows

$$L_{contras} = 1 \cdot \| E(x^1) - E(x^2) \|_2 + (1 - l) \cdot \max(0, m - \| E(x^1) - E(x^2) \|_2),$$  \hspace{1cm} (7)

where $x^1, x^2$ are two profile face images input to the dual-path architecture. The label vector $l$ indicates whether they contain the same ($l = 1$) or different ($l = 0$) identity. The hyper-parameter $m = 1.2$ indicates the margin.

**Total Objective:** In order to reduce the artifacts in the synthetic images, a total variation regularization $L_{tv}$ is also applied as that in [24]. The overall loss function used to train the proposed heterogeneous frontalization network is as follows

$$L = L_{front} + \lambda_C \cdot L_{contras} + \lambda_{cls} \cdot L_{cls},$$  \hspace{1cm} (8)

where $\lambda_C$, $\lambda_{cls}$ are regularization parameters.

**IV. EXPERIMENTS**

**Datasets:** We evaluate the proposed heterogeneous face frontalization network on three publicly available datasets: DEVCOM Army Research Laboratory Visible-Thermal Face Dataset (ARL-VTF) [37], ARL Multimodal Face Database (ARL-MMFD) [9], [17], [47] and TUFTS Face Dataset [33]. Sample images from these datasets are shown in Fig. 4.

**ARL-VTF [37] ARL-MMFD [9] TUFTS [33]**

Fig. 4. Input profile thermal images sampled from three datasets respectively.

**ARL-VTF Dataset:** The ARL-VTF dataset [37] consists of 500,000 images from 395 subjects (295 for training, 100 for testing). This dataset contains a large collection of paired visible and thermal faces. Variations in baseline, expressions, pose, and eye-wear are included. We first crop the images based on the given ground-truth bounding-box annotations for conducting experiments. The proposed model is trained on this dataset by the predefined development and testing splits [37]. The final experimental results are based on the average of the predefined 5 splits.

**ARL-MMFD Dataset:** Additionally, we evaluate the proposed model on Volume III of the ARL-MMFD dataset [9]. This dataset was collected by ARL cross 11 sessions over 6 days. It contains 5419 paired polarimetric thermal and visible images from 121 subjects with significant
variations in expression, illuminations, pose, glasses, etc. We select the polarimetric thermal profile and visible frontal face image pairs corresponding to both neutral and expressive faces for conducting experiments on this dataset. In particular, we randomly select images from 90 identities for training and the images from the remaining identities for testing. So, there are no overlapping images between training and testing sets. We use the original aligned visible and polarimetric thermal images for training and testing without any pre-processing. The final experimental results are based on the average of five random splits.

**TUFTS Face Dataset:** Finally, the proposed model is also evaluated on a recently published TUFTS Face Dataset [33]. TUFTS is a multi-modal dataset which contains more than 10,000 images from 113 individuals from more than 15 countries. For conducting experiments, we only select thermal and visible images from this multi-modal dataset. Hence, there are more than 1000 images from over 100 subjects with different pose and expression variations. This dataset is very challenging due to a large number of pose and expression variations and only a few images per variation are available in the dataset. Images from randomly selected 89 individuals are used for training and images from the remaining 23 subjects are used for testing. The raw images are used to train and test without any pre-processing. In particular, profile images in the thermal domain and the frontal images (both neural and expression) in the visible domain are utilized for training the models.

**Implementation Details:** All the images in this work are resized to $128 \times 128$ and the image intensity is scaled into $[0, 1]$. The features from VGGFace [35] average pooling layer are extracted from the synthesized visible image and the similarity for verification is calculated based on the cosine distance. In all the experiments, the hyper-parameters are set as $\lambda_I = 10$, $\lambda_{adv} = 1$, $\lambda_C = 0.01$, and $\lambda_{tv} = 1e^{-4}$. The learning rate is initially set as 0.01 and the batch size is 8. We train our model 10, 100, and 400 epochs on ARL-VTF, ARL-MMFD and TUFTS datasets, respectively.

We compare the proposed method with the following state-of-the-art facial frontalization methods: TP-GAN [19]; PIM [51]; M2FPA [28]; DA-GAN [44]. Besides, we add pix2pix [21] as a baseline, which is an image-to-image translation method between two domains. We followed the same network settings as mentioned in the original papers and tried our best to finetune the training parameters. Note that TP-GAN [19] and PIM [51] require landmarks on input profile thermal face images. We manually label the landmarks on thermal faces in the TUFTS Face Dataset [33] and use officially provided landmarks in the ARL-VTF dataset [37]. We estimate the landmarks by MTCNN [48] on the images in the ARL-MMFD dataset [9].

**A. Experimental Results**

**ARL Visible-Thermal Face Dataset:** The evaluation is based on the following two protocols: (1) Gallery G$_{VBO}$-to-Probe P$_{PTP}$. (2) Gallery G$_{VBO}$-to-Probe P$_{PTP}$. The images in Gallery G$_{VBO}$ are the facial baseline images in the thermal domain without eye-glasses occlusion. The images in Probe P$_{PTP}$ are the facial profile images in the thermal domain without eye-glasses occlusion. The images in Probe P$_{PTP}$- are the facial profile images in the thermal domain with eye-glasses token-off.

We evaluate our model with the other baseline methods both qualitatively and quantitatively. Visual frontalization results are shown in Fig. 5. As can be seen from this figure, the proposed method is able to synthesize more photo-realistic and identity-preserving images compared with the other methods. Other methods fail to synthesize high-quality and identity-preserving images. Additionally, Table IV quantitatively compares the verification performance of different methods. Our method achieves around 2% and 5% improvements on both the AUC and EER scores in two protocols, respectively when compared with the recent DA-GAN [44]. Additionally, the proposed model also achieves 2% ~ 5% improvements on the True Accept Rate (TAR) at False Accept Rates (FAR) 1% and 5%.

**ARL Multimodal Face Database:** We evaluate the proposed model on Volume III of ARL-MMFD [9]. Qualitative and quantitative results corresponding to this dataset are shown in Fig. 6 and Table III respectively. As can be seen from Table III, the proposed model surpasses the best performing baseline model by 2.4% and 1.5% in AUC and EER scores, respectively. Furthermore, the proposed model surpasses 7.7% and 1.3% when compared with the DA-GAN [43] at FAR=1% and FAR=5%, respectively.

From Fig. 6, we can see that our model is able to synthesize more photo-realistic images while preserving the identity better than the other frontalization models. For
TABLE I

VERIFICATION PERFORMANCE COMPARISON ON THE ARL-VTF DATASET [37].

| Gallery | Metric | AUC | EER | FAR=1% | FAR=5% |
|---------|--------|-----|-----|--------|--------|
| 0010    | Raw    | 54.88 | 46.38 | 2.16   | 8.33   |
|         | Ptx2Pix [21] | 5.04 | 46.84 | 2.17   | 8.90   |
|         | TP-GAN [19] | 64.21 | 40.12 | 3.31   | 12.11  |
|         | PIM [51] | 68.69 | 36.58 | 5.43   | 16.56  |
|         | M2FPA [28] | 74.99 | 32.33 | 5.73   | 20.26  |
|         | DA-GAN [44] | 75.58 | 31.18 | 6.85   | 22.23  |
|         | Ours    | 77.48 | 29.08 | 8.20   | 25.89  |

| Gallery | Metric | AUC | EER | FAR=1% | FAR=5% |
|---------|--------|-----|-----|--------|--------|
| 0010    | Raw    | 56.10 | 45.98 | 1.06   | 8.53   |
|         | Ptx2Pix [21] | 57.95 | 44.84 | 1.79   | 14.47  |
|         | TP-GAN [19] | 67.41 | 36.78 | 3.89   | 13.84  |
|         | PIM [51] | 73.31 | 32.70 | 5.96   | 20.42  |
|         | M2FPA [28] | 76.99 | 30.69 | 8.40   | 23.62  |
|         | DA-GAN [44] | 82.18 | 25.11 | 10.82  | 30.64  |

TABLE II

VERIFICATION PERFORMANCE COMPARISON CORRESPONDING TO THE ABLATION STUDY.

| Probe | Metric | AUC | EER | FAR=1% | FAR=5% |
|-------|--------|-----|-----|--------|--------|
| 00 pose | Baseline | 56.21 | 45.99 | 2.07   | 9.04   |
|       | w/ Multi-scale $L_1$ | 58.07 | 44.82 | 2.13   | 10.27  |
|       | w/ $L_{id}$ | 70.26 | 34.26 | 4.90   | 16.26  |
|       | w/ self-attn | 71.99 | 34.05 | 5.22   | 19.72  |
|       | w/ $L_{adv}$ | 72.58 | 33.54 | 6.21   | 20.07  |
|       | w/ Eq. (1) | 77.13 | 29.16 | 6.41   | 21.35  |
|       | w/ $L_{clss}$ | 77.18 | 29.07 | 7.43   | 22.20  |
|       | w/ $L_{contrast}$ (ours) | 77.48 | 29.08 | 8.20   | 25.89  |

| Gallery | Metric | AUC | EER | FAR=1% | FAR=5% |
|---------|--------|-----|-----|--------|--------|
| 10 pose | Baseline | 59.20 | 43.13 | 2.73   | 9.71   |
|         | w/ Multi-scale $L_1$ | 62.01 | 40.91 | 3.84   | 11.06  |
|         | w/ $L_{id}$ | 76.56 | 29.67 | 6.70   | 23.84  |
|         | w/ self-attn | 76.56 | 29.67 | 6.70   | 23.84  |
|         | w/ $L_{adv}$ | 76.58 | 28.44 | 6.66   | 24.18  |
|         | w/ Eq. (1) | 76.58 | 28.44 | 6.66   | 24.18  |
|         | w/ $L_{clss}$ | 76.56 | 28.44 | 6.66   | 24.18  |
|         | w/ $L_{contrast}$ (ours) | 80.56 | 26.46 | 9.24   | 29.90  |

TABLE III

VERIFICATION PERFORMANCE COMPARISON ON THE ARL-MMFD DATASET [9].

| Method  | AUC | EER | FAR=1% | FAR=5% |
|---------|-----|-----|--------|--------|
| Raw     | 64.12 | 40.51 | 4.21   | 17.74  |
| Ptx2Pix [21] | 73.60 | 31.96 | 11.16  | 26.45  |
| TP-GAN [19] | 76.15 | 30.89 | 6.26   | 19.03  |
| PIM [51] | 80.89 | 26.86 | 9.83   | 27.11  |
| M2FPA [28] | 85.58 | 22.27 | 13.32  | 37.58  |
| DA-GAN [44] | 86.26 | 21.56 | 14.74  | 33.17  |
| Ours    | 88.61 | 20.21 | 16.07  | 40.93  |

TABLE IV

VERIFICATION PERFORMANCE COMPARISON ON THE TUFTS FACE DATABASE [33].

| Method  | AUC | EER | FAR=1% | FAR=5% |
|---------|-----|-----|--------|--------|
| Raw     | 67.55 | 37.88 | 3.11   | 16.11  |
| Ptx2Pix [21] | 69.71 | 35.31 | 5.44   | 21.66  |
| TP-GAN [19] | 70.93 | 35.32 | 6.46   | 18.77  |
| PIM [51] | 72.84 | 34.10 | 8.77   | 21.00  |
| M2FPA [28] | 75.07 | 31.22 | 8.33   | 23.44  |
| DA-GAN [44] | 75.24 | 31.14 | 10.44  | 26.22  |
| Ours    | 78.68 | 28.38 | 10.44  | 27.11  |

Fig. 6. Cross-domain face frontalization comparison on the ARL-MMFD [9] dataset.

each subject in this dataset, the pose variations mainly cover from $-60^\circ$ to $60^\circ$, while the illumination variation is various. Hence, the method like PIM [51] fails to synthesize photo-realistic images due to the illumination artifacts. In addition, our proposed model is trained based on contrastive loss which helps the network provide better results even on this smaller dataset.

**TUFTS Face Database:** Finally, the proposed model is also evaluated on the TUFTS Face dataset [33]. The performances are reported in quantitative as shown in Table IV. We can observe our proposed method surpasses the previous baselines and achieves around 3% improvement in AUC and EER scores.

**B. Ablation Study**

In this section, we analyze how each part of the proposed model contributes to the final performance. We choose the global UNet [19] generator and the discriminator $D_g$ as the baseline model. In particular, we analyze the contribution of each loss function and the equalization in Eq (1). We conduct these ablation experiments on the ARL-VTF dataset and show the results both qualitatively and quantitatively in...
Fig. 7. Results corresponding to the ablation study.

As shown in Table II, the quantitative verification results are improved by consecutively adding different components. Fig. 7 shows the corresponding synthesized samples. Based on this ablation study, we can see that $L_{id}$ significantly preserves the identity in the synthesized image. Feature equalization Eq. (1) significantly improves the image quality. Finally, $L_{cls}$ and $L_{contra}$ improve the verification performance.

C. Pose Invariant Representation

In Fig. 9, we show the synthesized frontal faces from the proposed model with a range of yaw poses on the ARL-MMFD dataset. Yaw poses in this dataset mainly cover $0^\circ \sim 45^\circ$. Given arbitrary yaw pose polarimetric thermal images, we can observe that the synthesized visible frontal images maintain a good pose-invariant representation. Additionally, we also conduct this analysis on the ARL-VTF dataset as shown in Fig. 8. From these two figures, we can see that the proposed model learns pose-invariant frontal representation in the visible domain when given arbitrary pose thermal input images.

V. CONCLUSION

In this work, we proposed a novel heterogeneous face frontalization model which generates frontal visible faces from profile thermal faces. The generator contains a gradient reversal layer-based classifier for domain-agnostic feature learning and a pair of local and global discriminators for better synthesis. Additionally, another contrastive constraint is enforced by a dual-path training strategy for learning discriminative latent features. Quantitative and visual experiments conducted on three real thermal-visible datasets demonstrate the superiority of the proposed method when compared to other existing methods. Additionally, an ablation study was conducted to demonstrate the improvements obtained from different modules and loss functions.

REFERENCES

[1] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou. Large scale 3d morphable models. International Journal of Computer Vision, 126(2):233–254, 2018.
[2] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway. A 3d morphable model learnt from 10,000 faces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5543–5552, 2016.
[3] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun. Learning a high fidelity pose invariant model for high-resolution face frontalization. In NeurIPS, pages 2872–2882, 2018.
[4] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), volume 1, pages 539–546. IEEE, 2005.
[5] 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, 2019.
[6] X. Di and V. M. Patel. Facial synthesis from visual attributes via sketch using multiscale generators. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2(1):55–67, 2019.
[7] X. Di and V. M. Patel. Multimodal face synthesis from visual attributes. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021.
[8] 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), pages 1–8. IEEE, 2019.
[9] X. Di, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel. Multi-scale thermal to visible face verification via attribute guided synthesis. IEEE Transactions on Biometrics, Behavior, and Identity Science, 3(2):266–280, 2021.
[10] 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.
[11] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In International conference on machine learning, pages 1180–1189. PMLR, 2015.
