Explainable Thermal to Visible Face Recognition Using Latent-Guided Generative Adversarial Network

David Anghelone, Cunjian Chen, Philippe Faure, Arun Ross, Antitza Dantcheva

Explainable Thermal to Visible Face Recognition Using Latent-Guided Generative Adversarial Network

David Anghelone[1,2,3], Cunjian Chen[4], Philippe Faure[2], Arun Ross[4] and Antitza Dantcheva[1,3]
[1]Inria [2]Thales [3]Univsersité Côte d'Azur [4]Michigan State University

***Abstract*— One of the main challenges in performing thermal-to-visible face image translation is preserving the identity across different spectral bands. Existing work does not effectively disentangle the identity from other confounding factors. In this paper, we propose a Latent-Guided Generative Adversarial Network (LG-GAN) to explicitly decompose an input image into identity code that is spectral-invariant and style code that is spectral-dependent. By using such a disentanglement, we are able to analyze the identity preservation by interpreting and visualizing the identity code. We present extensive face recognition experiments on two challenging Visible-Thermal face datasets. We show that the learned identity code is effective in preserving the identity, thus offering useful insights on interpreting and explaining thermal-to-visible face image translation.**

I. Introduction

Face recognition beyond the visible spectrum allows for increased robustness in the presence of different *poses*, *illumination variations*, *noise*, as well as *occlusions*. Further benefits include incorporating the *absolute size of objects*, as well as *robustness to presentation attacks* such as makeup and masks. Therefore, comparing RGB face images against those acquired beyond the visible spectrum is of particular pertinence in designing Face Recognition (FR) systems for *defense, surveillance, and public safety* [7] and is referred to as Cross-spectral Face Recognition (CFR). While CFR brings the aforementioned benefits, it is more challenging than traditional FR for both *human examiners* as well as *computer vision algorithms*, due to following three limitations. Firstly, there can be large *intra-spectral variation*, where within the same spectrum, face samples of the same subject may exhibit larger variations in appearance than face samples of different subjects. Secondly, the appearance variation between two face samples of the same subject in different spectral bands can be larger than that of two samples belonging to two different subjects, referred to as *modality gap*. Finally, *limited* availability of *training samples* of cross-modality face image pairs can significantly impede learning-based schemes, including those based on deep learning models.

Thermal sensors have been widely deployed in nighttime and low-light environments for security and surveillance applications. Some of them capture face images beyond the visible spectrum. However, there is considerable performance degradation when a *direct* matching is performed between thermal (THM) face images and visible (VIS) face images (due to the modality gap). This is mainly due to the change in identity determining features across the thermal and visible domains.

Recent work on thermal-to-visible face recognition has predominantly focused on synthesizing visible faces from their thermal counterparts by generative adversarial networks, in order to minimize the spectral difference [18], [20], [17], [1], [4]. Cross-spectral identity matches were implicitly enforced by minimizing the reconstruction error [1], or by using an identity extraction network [18], [1], [4]. However, to the best of our knowledge, there is no prior thermal-to-visible generative adversarial network (GAN) that explicitly *encodes identity*.

Here, we propose a Latent-Guided Generative Adversarial Network (LG-GAN), which disentangles the latent space into an *identity* and a *style* space. In particular, latent space refers to the feature space learned by a GAN. While the identity space is shared by images of a subject acquired in different spectra, the style space is not. To translate an image from a source spectrum to a target spectrum, we combine the *identity code* with a *style code* denoting the target domain. Hence, the identity code is invariant across spectral domains, whereas the style code encodes the spectral information. To enable thermal-to-visible translation and vice versa, LG-GAN incorporates three networks per spectrum, (i) identity encoder, (ii) style encoder and (iii) decoder. Thermal-to-visible translation is performed by switching the spectrum of the encoder-decoder pairs with the opposite spectrum of the input image. To generate realistic samples, we incorporate a set of loss functions to further constrain the identity space. The overall architecture of LG-GAN is illustrated in Figure 1.

The main contributions of this work are the following.

- We propose a novel *supervised learning* framework for CFR that translates facial images from one spectrum to another, while preserving the identity. The framework is adapted from MUNIT [9].
- We introduce four loss functions that facilitate both image as well as latent reconstructions.
- We analyze the latent space, which is decomposed into a shared *identity* space and a spectrum-dependent *style* space, by visualizing the encoding using heatmaps.
- We evaluate the proposed framework on two benchmark multispectral face datasets and achieve promising results with respect to *visual quality*, as well as face recognition *matching scores*.

The rest of the paper is organized as follows. Section II reviews recent work on thermal face recognition involv-

The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

ing generative models. Section III describes the proposed LG-GAN, placing emphasis on the latent space and the loss functions. Section IV discusses the experiments and results pertaining to image synthesis, identity latent code understanding, as well as face recognition results on two multispectral face datasets. Section V concludes the work.

II. Related Work

Generative adversarial networks (GANs) have shown remarkable results in image-to-image translation. In the context of CFR, GANs can be used to translate between spectral domains. For example, a thermal face image can be *translated* to a visible[1] face image. Existing work has predominantly focused on *supervised* image-to-image translation for paired samples. Supervision can be leveraged by minimizing the difference in identity [18], [20], [1], semantic attributes [4] or facial shape [17], [1] between the synthesized visible face images and the target visible face images. Zhang et al. [18] computed an identity loss function by features extracted from a certain layer of a fine-tuned VGG model on visible face images. However, computing the identity loss using a *single* layer does not allow for the extraction of multi-scale features. Therefore, the authors obtained a lower similarity match score between synthesized visible faces and target visible faces. To overcome this, Chen et al. [1] firstly trained a face recognition network based on the VGG-19 network from a large-scale face recognition dataset. Then, intermediate features from *multiple* layers of VGG-19 were concatenated, in order to facilitate the extraction of identity-specific features. Though identity information can be largely preserved via the use of the identity loss function, it has been observed that facial attributes or shape information had noticeable artifacts or distortions caused by the translation. Therefore, Di et al. [4] extracted a set of attributes from a pre-trained attribute prediction network, in order to synthesize attribute-preserved visible images from thermal counterparts. Wang et al. [17] introduced a facial landmark detector to capture the facial structures that are essential to the preservation of identity features.

Apart from formulating new loss functions, some researchers have designed novel network architectures for the generator such as a densely connected encoder-decoder structure [10], cascaded-in-cascaded blocks [12], and self-attention blocks [3], that enable generating higher quality images. Iranmanesh et al. [10] presented a coupled generative adversarial network (CpGAN) architecture that incorporated a densely connected encoder-decoder structure in the generator. Kezebou et al. [12] proposed to reuse features from earlier convolutional layers via a UNET-like architecture with cascaded-in-cascaded blocks. Di et al. [3] enhanced a GAN with self-attention modules to enable attention-guided image synthesis.

III. Proposed Network

We propose LG-GAN, a latent-guided generative adversarial network, designed for paired thermal-to-visible translation. Specifically, LG-GAN learns the *content* as well as the *style* pertaining to a face that we refer to as *identity* and *style* code, respectively, in the latent space. LG-GAN is inspired from MUNIT [9]. In LG-GAN, *identity* and *style* are essential in translating images from an input thermal domain to an output visible domain, thereby bridging the domain gap through (a) enforcing both image-level and latent-level reconstructions, and (b) supervising thermal-to-visible image translation with an identity preserving loss function. Note that MUNIT [9] does not deal with the problem of CFR.

A. Baseline Model

The generator comprises of three networks for each domain, viz., *Identity-Encoder, Style-Encoder* and *Decoder*, targeted to extract a domain-shared identity latent code and a spectrum-specific style latent code. The translated image is reconstructed by combining the identity code with the style code of the target spectrum. Figure 2 illustrates the auto-encoder architecture. In the discriminator, we adopt the multi-scale discriminator which enables generation of realistic images with refined details.

B. Formalization

Let $\mathcal{V}$ and $\mathcal{T}$ be the visible and thermal domains. Let $x_{vis} \in \mathcal{V}$ and $x_{thm} \in \mathcal{T}$ be drawn from the marginal distributions $x_{vis} \sim p_{\mathcal{V}}$ and $x_{thm} \sim p_{\mathcal{T}}$, respectively. Thermal-to-visible face recognition based on GAN-synthesis aims to estimate the conditional distribution $p_{\mathcal{V}|\mathcal{T}}(x_{vis}|x_{thm})$, where,

$$p_{\mathcal{V}|\mathcal{T}}(x_{vis}|x_{thm}) = \frac{p_{\mathcal{V},\mathcal{T}}(x_{vis}, x_{thm})}{p_{\mathcal{T}}(x_{thm})} \tag{1}$$

involves the joint distribution $p_{\mathcal{V},\mathcal{T}}(x_{vis}, x_{thm})$. As the joint distribution is not known, we adopt the assumption of "partially shared latent space" from MUNIT [9] as follows.

A pair $(x_{vis}, x_{thm}) \sim p_{\mathcal{V},\mathcal{T}}$ of images, corresponding to the same face from the joint distribution, can be generated through the support of

(a) the **identity latent code** $id \in \mathcal{I}$, which is shared by both domains (we also introduce the notation $id_{vis}, id_{thm} \in \mathcal{I}$ for better domain-identity formalization),

(b) the **style latent code** $s_m \in \mathcal{S}_{\mathcal{M}}$, where $(m, \mathcal{M}) \in \{(vis, \mathcal{V}), (thm, \mathcal{T})\}$, which is specific to the individual domain.

Hence, we proceed to approximate the joint distribution via the latent space of the following two phases.

Within-domain reconstruction phase

Firstly, the identity latent code and style latent code are extracted from the input images x_{vis} and x_{thm}:

$$E_{\mathcal{V}}(x_{vis}) = (id_{vis}, s_{vis}) \text{ and } E_{\mathcal{T}}(x_{thm}) = (id_{thm}, s_{thm}). \tag{2}$$

Then, given the embedding of Equation (2), the face is reconstructed via the generator,

$$G_{\mathcal{V}}(id_{vis}, s_{vis}) = x_{vis}^{rec} \text{ and } G_{\mathcal{T}}(id_{thm}, s_{thm}) = x_{thm}^{rec}, \tag{3}$$

[1]We use the term *visible* to suggest *visible spectrum*

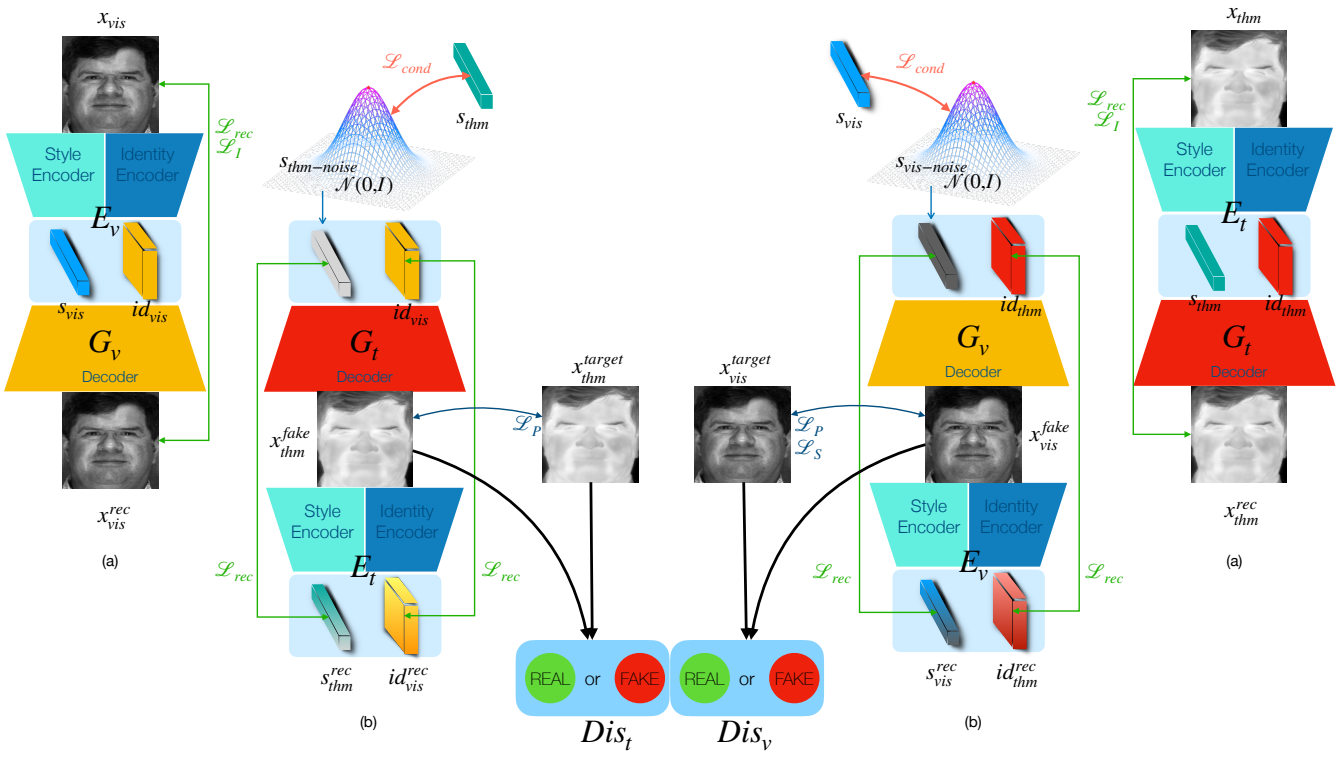

Fig. 1. Flowchart depicting the training of the proposed LG-GAN framework. It consists of two auto-encoders (E_v, G_v) and (E_t, G_t) dedicated to the visible and thermal domains, respectively. The sub-network (a) aims to learn the image reconstruction, while (b) enforcing the latent space reconstruction.

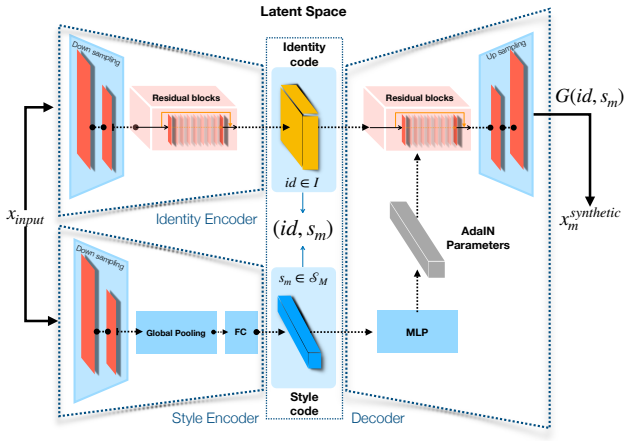

Fig. 2. The auto-encoder architecture incorporates three networks. An *identity encoder*, which extracts a domain-shared face identity latent code id from x_{input} and consists of convolutional layers followed by several residual blocks. A *style encoder*, which extracts a domain-specific spectral information latent code s_m from x_{input} and consists of convolutional layers followed by a global average pooling layer and a last fully connected layer. A *decoder*, which reconstructs the image from prior identity and style code (id, s_m) generating $x_m^{synthetic}$.

in order to learn the latent space for the specific face. Here, $\mathcal{M} \in \{\mathcal{V}, \mathcal{T}\}$ represents the domain, $E_\mathcal{M}$ denotes the factorized identity code and style code auto-encoder, $G_\mathcal{M}$ is the underlying decoder, and x_{vis}^{rec} and x_{thm}^{rec} are the corresponding reconstructed images.

The objective of LG-GAN is to learn the global *image reconstruction* mapping for a fixed $m \in \{vis, thm\}$, i.e.,

$$x_m \rightarrow x_m^{rec}, \tag{4}$$

while preserving facial identity features and allowing for a non-identity shift through *latent reconstruction* between

$$id_m \rightarrow id_m^{rec} \quad \text{and} \quad s_m \rightarrow s_m^{rec} \tag{5}$$

and forcing

$$s_{m-noise} \rightarrow s_m, \tag{6}$$

where, (id_m^{rec}, s_m^{rec}) are part of the extraction $(E_{\bar{\mathcal{M}}}(G_{\bar{\mathcal{M}}}(id_m, s_{\bar{m}-noise})), E_\mathcal{M}(G_\mathcal{M}(id_{\bar{m}}, s_{m-noise})))$, respectively, and $s_{m-noise}$ is randomly drawn from a prior normal distribution in order to learn the associated style distribution. $\bar{\mathcal{M}}$ and $\bar{m}$ represent opposite domains.

Cross-domain translation phase

In the domain translation phase, image-to-image translation is performed by swapping the encoder-modality (i.e., spectrum) with the opposite modality of the input image and imposing an explicit supervision on the style domain transfer functions $E_\mathcal{V}(x_{thm}) = (id_{thm}, s_{vis-noise})$ and $E_\mathcal{T}(x_{vis}) = (id_{vis}, s_{thm-noise})$, and then using $G_\mathcal{V}(id_{thm}, s_{vis-noise})$ and $G_\mathcal{T}(id_{vis}, s_{thm-noise})$ to produce the final output image $x_{vis/thm}^{fake}$ in the target spectrum. This is formalized as follows.

$$\begin{array}{rccc} \Theta_{t\rightarrow v} & : & \mathcal{T} & \rightarrow & \mathcal{V} \\ & & x_{thm} & \mapsto & x_{vis}^{fake} = G_\mathcal{V}(E_\mathcal{V}(x_{thm})); \end{array} \tag{7}$$

$$\begin{array}{rccc} \Theta_{v\to t} & : & \mathcal{V} & \to & \mathcal{T} \\ & & x_{vis} & \mapsto & x_{thm}^{fake} = G_{\mathcal{T}}(E_{\mathcal{T}}(x_{vis})). \end{array} \tag{8}$$

Consequently, $\Theta_{t\to v}$ and $\Theta_{v\to t}$ are the functions that synthesize the corresponding visible ($t \to v$) and thermal ($v \to t$) faces. Finally, LG-GAN learns the spectral conditional distribution $p_{\mathcal{V}|\mathcal{T}}(x_{vis}^{fake}|x_{thm})$ and $p_{\mathcal{T}|\mathcal{V}}(x_{thm}^{fake}|x_{vis})$ through a guided latent generation, where both these conditional distributions overcome the fact that we do not have access to the joint distribution $p_{\mathcal{V},\mathcal{T}}(x_{vis}, x_{thm})$. Indeed, the method is able to generate, as an alternative, the joint distributions $p_{\mathcal{V},\mathcal{T}}(x_{vis}^{rec}, x_{thm}^{fake})$ and $p_{\mathcal{V},\mathcal{T}}(x_{vis}^{fake}, x_{thm}^{rec})$, respectively. We aim to learn the translation using neural networks, and this paper will focus on Equation (7), where thermal face images are translated into realistic synthetic visible face images.

C. Loss Functions

LG-GAN is trained with the help of objective functions that include adversarial and bi-directional reconstruction loss as well as conditional, perceptual, identity, and semantic loss. We investigate the impact of each loss with respect to visual results and then propose an efficient combination. Further, we use the VGG-19 [14] architecture which, when trained on a specific dataset, could be used to extract relevant features prior to applying the loss functions.

1) Adversarial Loss: Images generated during the translated phase through Equations (7) and (8) must be realistic and not distinguishable from real images in the target domain. Therefore, the objective of the generators, Θ, is to maximize the probability of the discriminator $\mathbf{Dis}$ making incorrect decisions. The objective of the discriminator $\mathbf{Dis}$, on the other hand, is to maximize the probability of making a correct decision, i.e., to effectively distinguish between real and fake (synthesized) images.

$$\begin{aligned}\mathcal{L}_{GAN}^{t\to v} = \mathbb{E}_{x_{vis}\sim p_{\mathcal{V}}}[log(\mathbf{Dis}_{\mathcal{V}}(x_{vis}))] + \\ \mathbb{E}_{x_{thm}\sim p_{\mathcal{T}}}[log(1-\mathbf{Dis}_{\mathcal{V}}(\Theta_{t\to v}(x_{thm})))],\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{GAN}^{v\to t} = \mathbb{E}_{x_{thm}\sim p_{\mathcal{T}}}[log(\mathbf{Dis}_{\mathcal{T}}(x_{thm}))] + \\ \mathbb{E}_{x_{vis}\sim p_{\mathcal{V}}}[log(1-\mathbf{Dis}_{\mathcal{T}}(\Theta_{v\to t}(x_{vis})))].\end{aligned}$$

The adversarial loss is denoted as follows.

$$\mathcal{L}_{GAN} = \mathcal{L}_{GAN}^{t\to v} + \mathcal{L}_{GAN}^{v\to t}. \tag{9}$$

2) Bi-directional Reconstruction Loss: Loss functions in the Encoder-Decoder network encourage the domain reconstruction with regards to both the image reconstruction and latent space (identity+style) reconstruction.

$$\mathcal{L}_{rec}^{image} = \mathop{\mathbb{E}}_{x_m^{rec}; x_m\sim p_{\mathcal{M}}}[\|x_{vis}^{rec} - x_{vis}\|_1 + \|x_{thm}^{rec} - x_{thm}\|_1], \tag{10}$$

$$\mathcal{L}_{rec}^{identity} = \mathop{\mathbb{E}}_{id_m^{rec}; id_m\sim p_{\mathcal{M}}}[\|id_{vis}^{rec} - id_{vis}\|_1 + \|id_{thm}^{rec} - id_{thm}\|_1], \tag{11}$$

$$\mathcal{L}_{rec}^{style} = \mathop{\mathbb{E}}_{s_m^{rec}; s_m\sim\mathcal{N}}[\|s_{vis}^{rec} - s_{vis}\|_1 + \|s_{thm}^{rec} - s_{thm}\|_1]. \tag{12}$$

The bi-directional[2] reconstruction loss function is computed as follows:

$$\mathcal{L}_{rec} = \mathcal{L}_{rec}^{image} + \mathcal{L}_{rec}^{identity} + \mathcal{L}_{rec}^{style}. \tag{13}$$

3) Conditional Loss: Imposing a condition on the spectral distribution is essential and is the major difference from the baseline model [9]. Indeed, this allows for a translation that is conditioned the distribution of the target style code and, further, adds an explicit supervision on the final mapping $\Theta_{t\to v}$ and $\Theta_{v\to t}$. The conditional loss $\mathcal{L}_{cond}$ is defined as follows.

$$\begin{aligned}\mathcal{L}_{cond} = \mathop{\mathbb{E}}_{s_{vis-noise}; s_{vis}\sim\mathcal{N}}\|s_{vis-noise} - s_{vis}\|_1 \\ + \mathop{\mathbb{E}}_{s_{thm-noise}; s_{thm}\sim\mathcal{N}}\|s_{thm-noise} - s_{thm}\|_1.\end{aligned} \tag{14}$$

To improve the quality of the synthesized images and render them more realistic, we incorporate three additional objective functions.

4) Perceptual Loss: The perceptual loss $\mathcal{L}_P$ affects the perceptive rendering of the image by measuring the high-level semantic difference between synthesized and target face images. It reduces artefacts and enables the reproduction of realistic details [11]. $\mathcal{L}_P$ is defined as follows:

$$\begin{aligned}\mathcal{L}_P = \mathop{\mathbb{E}}_{x_{vis}^{fake}; x_{vis}\sim p_{\mathcal{V}}}\|\phi_P(x_{vis}^{fake}) - \phi_P(x_{vis})\|_1 \\ + \mathop{\mathbb{E}}_{x_{thm}^{fake}; x_{thm}\sim p_{\mathcal{T}}}\|\phi_P(x_{thm}^{fake}) - \phi_P(x_{thm})\|_1,\end{aligned} \tag{15}$$

where, ϕ_P represents features extracted by VGG-19, pre-trained on ImageNet.

5) Identity Loss: The identity loss $\mathcal{L}_I$ is responsible for preserving identity-specific features during the image reconstruction phase and, therefore, encourages the translated image to preserve the identity content of the input image. $\mathcal{L}_I$ is defined as follows:

$$\begin{aligned}\mathcal{L}_I = \mathop{\mathbb{E}}_{x_{vis}^{rec}; x_{vis}\sim p_{\mathcal{V}}}\|\phi_I(x_{vis}^{rec}) - \phi_I(x_{vis})\|_1 \\ + \mathop{\mathbb{E}}_{x_{thm}^{rec}; x_{thm}\sim p_{\mathcal{T}}}\|\phi_I(x_{thm}^{rec}) - \phi_I(x_{thm})\|_1,\end{aligned} \tag{16}$$

where, ϕ_I denotes the features extracted from the VGG-19 network pre-trained on the large-scale VGGFace2 dataset.

6) Semantic Loss: The semantic loss $\mathcal{L}_S$ guides the texture synthesis from thermal to visible domain and imparts attention to specific facial details. A parsing network is used to detect semantic labels and to classify them into 19 different classes which correspond to the segmentation mask of facial attributes provided by CelebAMask-HQ [13]. We apply semantic face parsing to images in our datasets. A few examples are shown in Figure 3. $\mathcal{L}_S$ is defined as follows.

$$\begin{aligned}\mathcal{L}_S = \mathop{\mathbb{E}}_{x_{vis}^{fake}; x_{vis}\sim p_{\mathcal{V}}}\|\phi_S(x_{vis}^{fake}) - \phi_S(x_{vis})\|_1 \\ \mathop{\mathbb{E}}_{x_{thm}^{fake}; x_{thm}\sim p_{\mathcal{T}}}\|\phi_S(x_{thm}^{fake}) - \phi_S(x_{thm})\|_1,\end{aligned} \tag{17}$$

where, ϕ_S is the parsing network, providing corresponding parsing class label.

[2]Bi-directional refers to the reconstruction learning process between *image* $\to$ *latent* $\to$ *image* and *latent* $\to$ *image* $\to$ *latent* by the sub-network (a) and (b), respectively, depicted in Figure 1.

Fig. 3. Example of face parsing results guided by the 19-class semantic label, when applied to images in the ARL-MMFD [8] and ARL-VTF [15] datasets.

Total loss: The overall loss function for the proposed LG-GAN is denoted as follows:

$$\min_{E_{\mathcal{V}},E_{\mathcal{T}},G_{\mathcal{V}},G_{\mathcal{T}}} \max_{\mathbf{Dis}} \mathcal{L}(E_{\mathcal{V}},E_{\mathcal{T}},G_{\mathcal{V}},G_{\mathcal{T}},\mathbf{Dis}) =$$

$$\lambda_{GAN}\mathcal{L}_{GAN}+\lambda_{rec}\mathcal{L}_{rec}+\lambda_{cond}\mathcal{L}_{cond}+\lambda_P\mathcal{L}_P+\lambda_I\mathcal{L}_I+\lambda_S\mathcal{L}_S. \quad (18)$$

Implementation Details

We implement the proposed LG-GAN framework in Pytorch by adapting MUNIT and designing the architecture for the modality-translation task. We note that we omit their proposed domain-invariant perceptual loss as well as the style-augmented cycle consistency. We train LG-GAN until convergence. The initial learning rate for Adam optimization is 0.0001 with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. For all experiments, the batch size is set to 1 and, based on empirical analysis, the loss weights are set to $\lambda_{GAN} = 1$, $\lambda_{rec} = 10$, $\lambda_{cond} = 35$, $\lambda_P = 15$, $\lambda_I = 20$ and $\lambda_S = 10$.

IV. Experimental Results

A. Dataset and Protocol

1) ARL-MMFD Dataset: The ARL-MultiModal Face dataset [8] (ARL-MMFD) contains visible, LWIR, and polarimetric face images of over 60 subjects and includes variations in both expression and standoff distances. We only use visible and LWIR (i.e., thermal) images for our experiment at one particular stand-off distance: 2.5m. The first 30 subjects are used for testing and evaluation, and the remaining 30 subjects are used for training. The images in this dataset are already aligned and cropped.

2) ARL-VTF dataset: The ARL-Visible Thermal Face dataset [15] (ARL-VTF) represents the largest collection of paired visible and thermal face images acquired in a time-synchronized manner. It contains data from 395 subjects with over 500,000 images captured with variations in expression, pose, and eyewear. We follow the established evaluation protocol, which assigns 295 subjects for training and 100 subjects for testing and evaluation. We select the baseline gallery and probe subjects without glasses, named *G VB0-* and *P TB0-*, respectively. Furthermore, we align and process the images based on the provided *eyes, nose* and *mouth* landmarks. Figure 4 depicts an example of such an alignment.

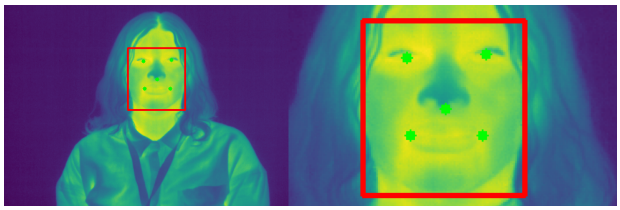

Fig. 4. Face alignment and cropping based on *eyes, nose* and *mouth* landmarks applied on images in the ARL-VTF dataset. [15]

B. Face recognition Performance

1) Face recognition Matcher: The ArcFace matcher [2] was trained on normalized face images of size 112×112 from the MS-Celeb-1M dataset [6] with additive angular margin loss. ResNet-50 was used as the embedding network and the final embedded feature size was set to 512. l_2 normalization was applied to the extracted feature vectors (embeddings), prior to the computation of cosine distance (match) scores.

2) Evaluation on Datasets: LG-GAN aims to decompose the latent space. However, MUNIT, that serves as the basis for our method, performs image translation in an unsupervised manner and cannot be employed in our thermal-to-visible scenario, as facial identity would not be preserved. Therefore, we incorporate $\mathcal{L}_{cond}$ (Equation (14)) as a conditional constraint forcing latent reconstruction (Equation (6)) with a normal noise distribution. Thus, the MUNIT-like supervised-approach, denoted as $\mathcal{L}_{base}$, will serve as a reference baseline model in our study.

We conduct face verification experiments on the ARL-MMFD and ARL-VTF datasets. Area Under the Curve (AUC) and Equal Error Rate (EER) metrics are computed using the ArcFace matcher. Table I reports face verification results on the ARL-MMFD and ARL-VTF datasets. A higher AUC indicates better performance, whereas a lower EER is better. We observe that translating thermal face images into visible-like face images significantly boosts the verification performance. For example, the *direct comparison* approach, in which we compare a thermal probe directly to the visible gallery, is related to the lowest AUC and highest EER scores, viz., 73.71%AUC and 32.73%EER in ARL-MMFD, and 54.80%AUC and 46.36%EER in ARL-VTF. When we apply the $\mathcal{L}_{base}$ from the baseline approach, and incorporate the adversarial $\mathcal{L}_{GAN}$ (Equation (9)), bi-directional reconstruction $\mathcal{L}_{rec}$ (Equation (13)) and conditional $\mathcal{L}_{cond}$ (Equation (14)) loss functions, the performance improves to 79.33%AUC and 29.16%EER on ARL-MMFD, and 92.21%AUC and 15.88%EER on ARL-VTF. This confirms that image-to-image translation significantly reduces the modality gap. Our proposed LG-GAN that includes $\mathcal{L}_{P+I+S}$ (Equation (18)), built on the basis of $\mathcal{L}_{base}$ that is improved by adding the perceptual $\mathcal{L}_P$ (Equation (15)), identity $\mathcal{L}_I$ (Equation (16)) and semantic $\mathcal{L}_S$ (Equation (17)) loss functions, exhibits the best performance of 93.99%AUC and 13.02%EER on ARL-MMFD, and 94.26%AUC and 12.99%EER on ARL-VTF. In optimizing LG-GAN on the large-scale dataset ARL-VTF, we tune the hyper-parameters (weights), thereby

TABLE I

FACE VERIFICATION PERFORMANCE, IMAGE QUALITY, AND IMPACT OF DIFFERENT LOSS FUNCTIONS ON ARL-MMFD [8] AND ARL-VTF [15] DATASETS. $\mathcal{L}_{base}$ REPRESENTS THE METHOD INCLUDING THE ADVERSARIAL (9) BI-DIRECTION RECONSTRUCTION (13) AND CONDITIONAL (14) LOSSES, WHILE $\mathcal{L}_P$, $\mathcal{L}_I$, $\mathcal{L}_{P+I}$ AND $\mathcal{L}_{P+I+S}$ ARE THE PERCEPTUAL (15), IDENTITY (16) AND SEMANTIC (17) LOSSES ADDED TO THE ORIGINAL $\mathcal{L}_{base}$ TRAINING.

	ARL-MMFD Dataset [8]			**ARL-VTF Dataset** [15]		
	AUC (%)	EER (%)	SSIM	AUC (%)	EER (%)	SSIM
Direct comparison	73.71	32.73	0.2899	54.80	46.31	0.3739
$\mathcal{L}_{base}$	79.33	29.16	0.4409	92.21	15.88	0.6049
$\mathcal{L}_P$	86.99	21.09	0.4596	92.79	14.24	0.6129
$\mathcal{L}_I$	84.20	22.90	0.4549	92.98	13.01	0.6101
$\mathcal{L}_{P+I}$	87.63	19.40	0.4626	92.15	15.36	0.6136
$\mathcal{L}_{P+I+S}$ = *LG-GAN*	**93.99**	**13.02**	**0.4652**	94.26	12.99	0.6145
LG-GAN optimized				**96.96**	**5.94**	**0.6787**

enabling objective functions to be combined in an effective manner. We set $\lambda_{rec} = 20$ placing emphasis on both image reconstruction $\mathcal{L}_{rec}^{image}$ (Equation (10)) and identity code reconstruction $\mathcal{L}_{rec}^{identity}$ (Equation (11)). On the other hand, we decrease $\lambda_I = 10$ towards improving the face verification accuracy to 96.96%AUC and 5.94%EER.

3) Comparison with State-of-the-Art Methods: We proceed to compare the proposed LG-GAN with state-of-the-art GAN-based CFR. Table II and Table III summarize AUC and EER scores as reported by other authors on the ARL-MMFD and ARL-VTF datasets, respectively. LG-GAN outperforms all other methods on the ARL-MMFD dataset and is competitive with SAGAN.

TABLE II

COMPARISON OF LG-GAN WITH OTHER SYNTHESIS-BASED APPROACHES ON THE ARL-MMFD DATASET.

Method	AUC (%)	EER (%)
GAN-VFS [18]	79.30	27.34
AP-GAN [5]	84.16	23.90
AP-GAN (GT) [5]	86.08	23.13
Multi-stream GAN [19]	85.74	23.18
SAGAN [3]	91.49	15.45
SG-GAN [1]	93.08	14.24
Multi-AP-GAN [4]	90.74	18.20
Multi-AP-GAN (GT) [4]	92.72	16.05
LG-GAN (ours)	**93.99**	**13.02**

TABLE III

COMPARISON OF LG-GAN WITH GAN-BASED CFR METHODS ON ARL-VTF DATASET.

Method	AUC (%)	EER (%)
Pix2Pix	71.12	33.80
GAN-VFS [18]	97.94	8.14
SAGAN [3]	**99.28**	**3.97**
LG-GAN (ours)	94.26	12.99
LG-GAN optimized (ours)	96.96	5.94

C. Ablation study

To illustrate the impact of loss functions included in LG-GAN on visual quality, we conduct an ablation study using both ARL-MMFD and ARL-VTF datasets. We evaluate the quality of generated images by the structural similarity index measure (SSIM) [16], where an SSIM score of 1 is the extreme case of comparing identical images. Table I reports average SSIM scores computed on both datasets under different experimental configurations.

The first intuitive observation has to do with the low performance of direct matching between thermal and visible face images that can be explained in terms of the lower SSIM of 0.2899 and 0.3739, respectively, on the ARL-MMFD and ARL-VTF datasets. This illustrates once more the modality gap. In an attempt to overcome this modality gap, the first baseline experiment $\mathcal{L}_{base}$, without visual quality optimization, boosts the SSIM score to 0.4409 and 0.6049, respectively. However, we note that generated results are rather blurry. By adding additional loss functions to $\mathcal{L}_{base}$, $\mathcal{L}_P$ (Equation (15)) and $\mathcal{L}_I$ (Equation (16)), namely the perceptual and identity losses, the SSIM score only marginally increases. Jointly, $\mathcal{L}_{P+I}$ improves the SSIM score further. Finally, the proposed LG-GAN is able to generate more realistic images with less artefacts, and with higher similarity to the visible ground-truth images. The resulting SSIM scores are 0.4652 and 0.6145, respectively. When we optimize LG-GAN on the large scale VTF dataset, we observe an SSIM score of 0.6787. We show the synthesized visible images in Figure 5. In Figure 8, we illustrate with the help of SSIM similarity and difference scaling, the face regions that are most sensitive to particular loss functions. Besides the impact of individual and combined loss functions on the visual quality of images, we also demonstrate their related impact on the face verification performance. Figure 6 and Figure 7 depict ROC curves pertaining to different loss functions for the ARL-MMFD and ARL-VTF datasets, respectively. We observe the correlation between SSIM scores and CFR matching performance.

D. Latent code visualization

Understanding the latent code is critical for LG-GAN, as it aims to elicit identity-specific information while ignoring spectrum induced information. A disentangled latent space is produced using an identity encoder and a style encoder that decomposes the input image into an identity code and a style code; see Figure 2. As discussed earlier, the style code represents the spectral information and drives the

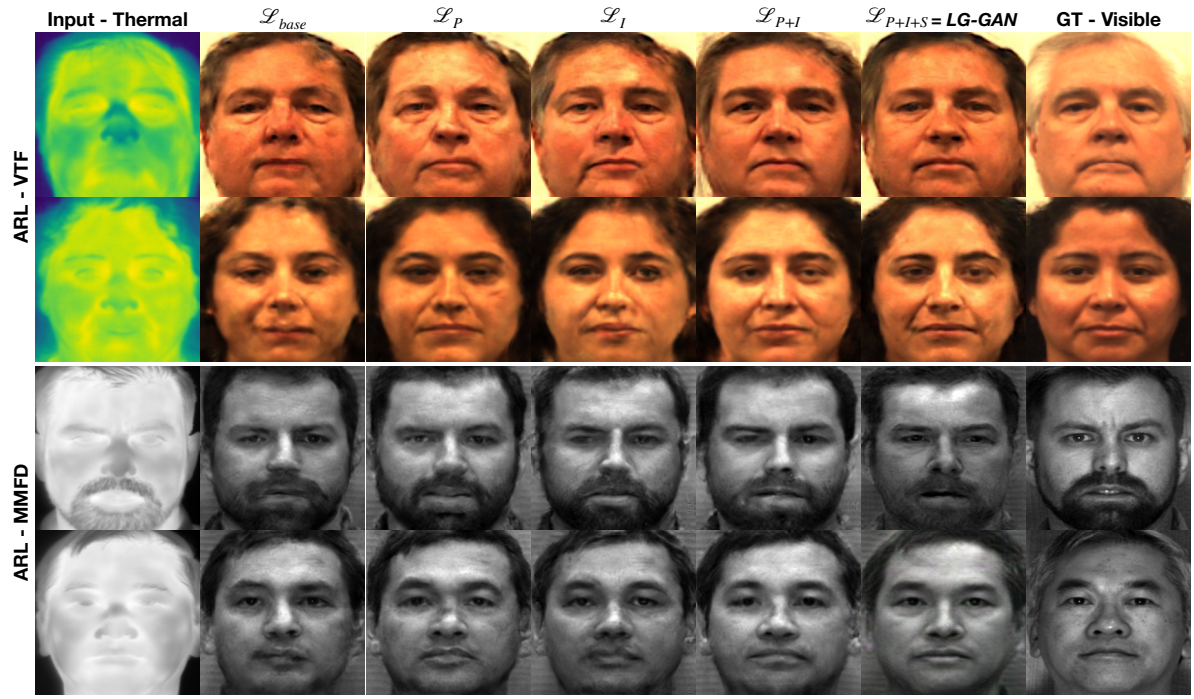

Fig. 5. Synthesizing visible face images from thermal images on the ARL-VTF [15] (top) and ARL-MMFD [8] (bottom) datasets using LG-GAN. This illustration also shows the impact of different loss functions and combinations thereof on the visual result. Synthesis using $\mathcal{L}_{base}$ includes the adversarial $\mathcal{L}_{GAN}$ (Eq.(9)), bi-directional reconstruction $\mathcal{L}_{rec}$ (Eq.(13)) and conditional $\mathcal{L}_{cond}$ (Eq.(14)) loss functions. $\mathcal{L}_P$, $\mathcal{L}_I$, $\mathcal{L}_{P+I}$ and $\mathcal{L}_{P+I+S}$ pertain to the addition of perceptual eq.(15), identity eq.(16) and semantic eq.(17) loss functions to the original $\mathcal{L}_{base}$ during the training stage.

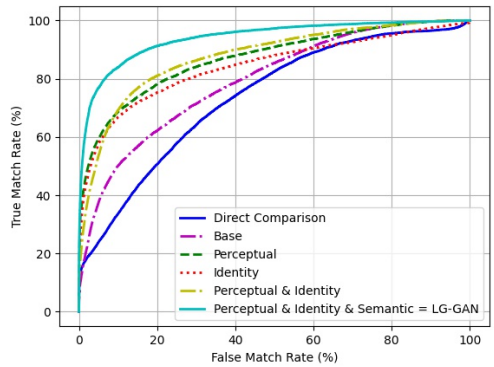

Fig. 6. ROC curves. An ablation study of different loss functions on the MMFD dataset. Our proposed LG-GAN achieves better performance than the baseline model, "Base".

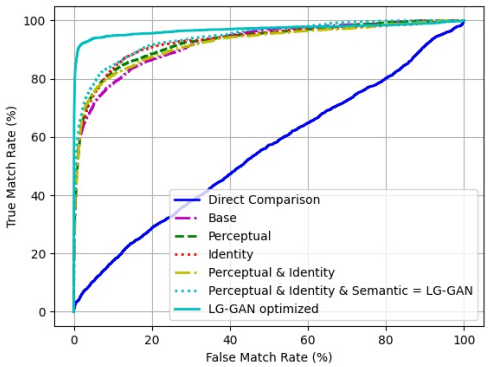

Fig. 7. ROC curves. An ablation study of different loss functions on the VTF dataset. Our proposed LG-GAN achieves better performance than the baseline model, "Base".

domain translation, without adversely affecting the identity information. However, here we seek to explore whether identity is explicitly encoded in the latent space. Towards this goal, we visualize the identity code, id_m, directly after the encoding step $E_{\mathcal{M}}(x_m)$; then, by up-scaling the code to the target image size, we determine the pertinent pixels that are responsible for the identity information in the latent space. This is visualized in Figure 9. We observe that facial features around eyes, nose, mouth and hair have been encoded. Moreover, identity codes – id_{vis} and id_{thm} – extracted from both spectra also highlight the same visual information. This is consistent with the partially shared latent space assumption made by Huang et al. [9].

V. Conclusions

In this paper, we propose a latent-guided generative adversarial network (LG-GAN) that explicitly decomposes an input image into an identity code and a style code. The identity code is learned to encode spectral-invariant identity features between thermal and visible image domains in a supervised setting. In addition, the identity code offers useful insights in explaining salient facial structures that are essential to the synthesis of high-fidelity visible spectrum face images.

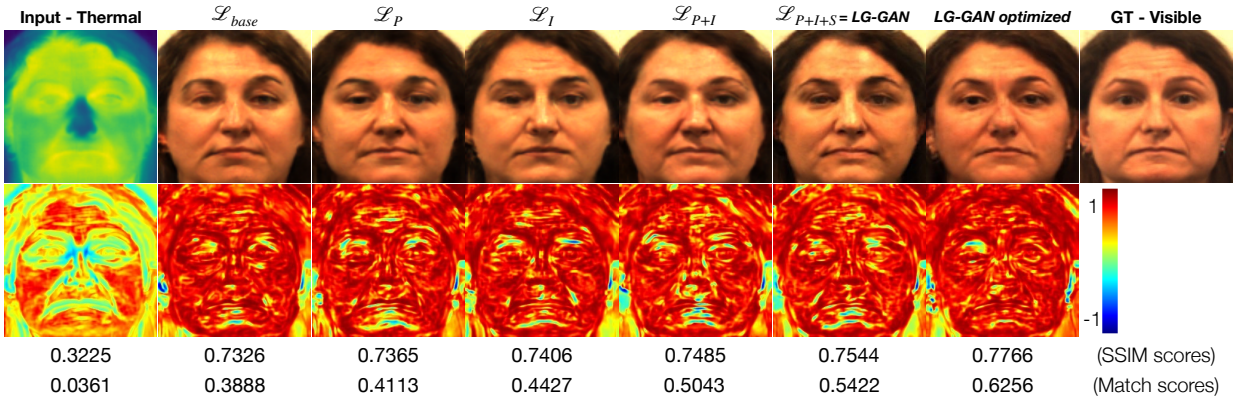

Fig. 8. A visualization strategy to understand the impact of different loss functions on the visual quality as well as the matching scores. The top row shows samples generated by individual and combined loss functions, while the bottom row illustrates the SSIM scores as well as the SSIM *similarity and difference* of two images in different scenarios: *GT-Visible* against *Input-Thermal/*$\mathcal{L}_{base}$*/*$\mathcal{L}_P$*/*$\mathcal{L}_I$*/*$\mathcal{L}_{P+I}$*/LG-GAN/LG-GAN optimized*.

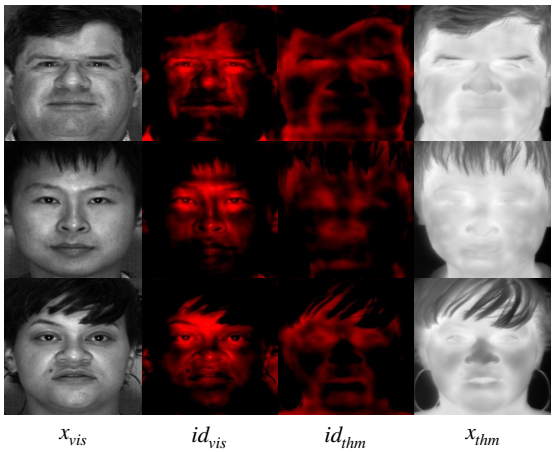

Fig. 9. Visualization of identity codes, id_{vis} and id_{thm}, extracted from $E_{\mathcal{V}}(x_{vis})$ and $E_{\mathcal{T}}(x_{thm})$, respectively.

Experiments on two datasets suggest that our proposed LG-GAN achieves competitive thermal-to-visible cross-spectral face recognition accuracy, while enabling explanations on salient features used for thermal-to-visible image translation. Future work will involve enhancing the identity code representation with attention modules and visualizing the style code, in order to deepen understanding of thermal-to-visible image translation.

REFERENCES

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

[2] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[3] 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 *International Conference on Biometrics*, pages 1–8, 2019.

[4] 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.

[5] X. Di, H. Zhang, and V. M. Patel. Polarimetric thermal to visible face verification via attribute preserved synthesis. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, 2018.

[6] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, 2016.

[7] S. Hu, N. Short, B. S. Riggan, M. Chasse, and M. S. Sarfraz. Heterogeneous face recognition: Recent advances in infrared-to-visible matching. In *IEEE International Conference on Automatic Face Gesture Recognition*, pages 883–890, 2017.

[8] S. Hu, N. J. Short, B. S. Riggan, C. Gordon, K. P. Gurton, M. Thielke, P. Gurram, and A. L. Chan. A polarimetric thermal database for face recognition research. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[9] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision*, 2018.

[10] S. M. Iranmanesh, B. Riggan, S. Hu, and N. M. Nasrabadi. Coupled generative adversarial network for heterogeneous face recognition. *Image and Vision Computing*, 94:103861, 2020.

[11] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *European Conference on Computer Vision*, pages 694–711, 2016.

[12] L. Kezebou, V. Oludare, K. Panetta, and S. Agaian. TR-GAN: thermal to RGB face synthesis with generative adversarial network for cross-modal face recognition. In *Mobile Multimedia/Image Processing, Security, and Applications*, volume 11399, pages 158 – 168, 2020.

[13] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[14] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 2015.

[15] D. Poster, M. Thielke, R. Nguyen, S. Rajaraman, X. Di, C. N. Fondje, V. M. Patel, N. J. Short, B. S. Riggan, N. M. Nasrabadi, and S. Hu. A large-scale, time-synchronized visible and thermal face dataset. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1559–1568, 2021.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, pages 600–612, 2004.

[17] Z. Wang, Z. Chen, and F. Wu. Thermal to visible facial image translation using generative adversarial networks. *IEEE Signal Processing Letters*, 25(8):1161–1165, 2018.

[18] H. Zhang, V. M. Patel, B. S. Riggan, and S. Hu. Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces. In *IEEE International Joint Conference on Biometrics*, pages 100–107, 2017.

[19] H. Zhang, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel. Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. *International Journal of Computer Vision*, 127(6):845–862, 2019.

[20] T. Zhang, A. Wiliem, S. Yang, and B. Lovell. Tv-gan: Generative adversarial network based thermal to visible face recognition. In *International Conference on Biometrics*, pages 174–181, 2018.