Towards a Robust Thermal-Visible Heterogeneous Face Recognition Approach Based on a Cycle Generative Adversarial Network

Nadir Kamel Benamara[1]*, Ehlem Zigh[2], Tarik Boudghene Stambouli[1], Mokhtar Keche[1]

[1] Laboratoire Signaux et Images, Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, USTO-MB, BP1505, El M'naouer, 31000, Oran (Algeria)
[2] Laboratoire LaRATIC, Institut National des Télécommunications et des TIC d'Oran, BP 1518, El M'nouer, 31000 Oran (Algeria)

Received 6 March 2021 | Accepted 18 October 2021 | Published 20 December 2021

Abstract

Security is a sensitive area that concerns all authorities around the world due to the emerging terrorism phenomenon. Contactless biometric technologies such as face recognition have grown in interest for their capacity to identify probe subjects without any human interaction. Since traditional face recognition systems use visible spectrum sensors, their performances decrease rapidly when some visible imaging phenomena occur, mainly illumination changes. Unlike the visible spectrum, Infrared spectra are invariant to light changes, which makes them an alternative solution for face recognition. However, in infrared, the textural information is lost. We aim, in this paper, to benefit from visible and thermal spectra by proposing a new heterogeneous face recognition approach. This approach includes four scientific contributions. The first one is the annotation of a thermal face database, which has been shared via Github with all the scientific community. The second is the proposition of a multi-sensors face detector model based on the last YOLO v3 architecture, able to detect simultaneously faces captured in visible and thermal images. The third contribution takes up the challenge of modality gap reduction between visible and thermal spectra, by applying a new structure of CycleGAN, called TV-CycleGAN, which aims to synthesize visible-like face images from thermal face images. This new thermal-visible synthesis method includes all extreme poses and facial expressions in color space. To show the efficacy and the robustness of the proposed TV-CycleGAN, experiments have been applied on three challenging benchmark databases, including different real-world scenarios: TUFTS and its aligned version, NVIE and PUJ. The qualitative evaluation shows that our method generates more realistic faces. The quantitative one demonstrates that the proposed TV-CycleGAN gives the best improvement on face recognition rates. Therefore, instead of applying a direct matching from thermal to visible images which allows a recognition rate of 47,06% for TUFTS Database, a proposed TV-CycleGAN ensures accuracy of 57,56% for the same database. It contributes to a rate enhancement of 29,16%, and 15,71% for NVIE and PUJ databases, respectively. It reaches an accuracy enhancement of 18,5% for the aligned TUFTS database. It also outperforms some recent state of the art methods in terms of F1-Score, AUC/EER and other evaluation metrics. Furthermore, it should be mentioned that the obtained visible synthesized face images using TV-CycleGAN method are very promising for thermal facial landmark detection as a fourth contribution of this paper.

Keywords

Deep Learning, Generative Adversarial Network, Heterogeneous Face Recognition, Thermal Sensor.

DOI: 10.9781/ijimai.2021.12.003

I. Introduction

Automatic identification has become a crucial routine task to recognize individuals efficiently without any human interaction, in many applications, such as e-payment [1], people tagging in social media, gaming,...etc. Based on invariant biological and/or behavior characteristics [2], like face, iris, palm, fingerprint signature or gait recognition, biometric technologies are mainly used for security, like access control and criminal identification.

* Corresponding author.
E-mail address: nadirkamel.benamara@univ-usto.dz

Face modality has many advantages. It is a natural recognition procedure, generally accepted by everyone. It is cheaper in comparison to iris or fingerprint recognition modalities, since the price of cameras has become more accessible with the massive development of CMOS/CCD sensors. It is easily done, non-intrusive and above all, it works at a distance and unobtrusively, which makes it suitable for highly populated places, such as airports or bus stations. Therefore, this modality has attracted several scientific researchers. We can cite for example [3]–[6] which have reached a significant level of recognition accuracy for controlled environments. Nevertheless, the performances of those systems, which use the visible light, with wavelength ranging from 0.4 μm to 0.8 μm, are highly dependent on lighting quality and intensity and cannot be used at all for night applications. Furthermore,

they face other challenges, such as pose and expression variations, and face disguises. To overcome these challenges, several alternative approaches have been proposed, they are mainly based on 3D or Infrared (IR) imagery.

With invariance to brightness changes as the main asset, IR imaging has emerged as a particularly promising research direction in the facial biometrics field [7]–[10]. It is a burgeoning sensor modality that could further be divided into two categories:

- **Active infrared**: it relies on signal reflected from objects illuminated by an infrared beam. It includes near-infrared (NIR) (0.74 μm - 1 μm) and short-wave infrared (SWIR) (1 μm - 3 μm).
- **Passive infrared**: it is based on body emitted radiation measurements, commonly known as thermal infrared. It comprises middle wave infrared (MWIR) (3 μm - 5 μm) and long-wave infrared (LWIR) (8 μm – 14 μm).

A Face Recognition (FR) system that aims to work under different illumination scenarios, during daytime as well as nighttime, should take into consideration both visible and infrared images. NIR and SWIR use an external active illuminator, which makes the active IR based FR systems improper for highly covert applications, since such an illuminator is easily detectable. Another major drawback for security applications is that NIR and SWIR images do not allow liveness verification natively, which makes the FR systems based on them, like those based on the visible images, vulnerable to spoofing attacks by just a photo or a video record (see Fig. 1). To overcome this issue, [11] has proposed eye-blink liveness checking. It achieves an interesting performance, but it was far away to be robust against the recent spoofing techniques, like 3D mask replica [12], [13]. Based just on body-emitted radiation, MWIR/LWIR remains an efficient solution against these issues (Fig. 2). It should be mentioned that few works have addressed the LWIR-VIS cross-spectral matching problem compared to the NIR-VIS problem, because the challenge intricacy increases proportionally with the spectrum wavelength. The authors in [14] have studied the modality gap based on the structural similarity index (SSIM) as a quantitative measure, and have obtained 0.335 for LWIR-VIS scenario and 0.581 for NIR-VIS scenario. Furthermore, public thermal LWIR face databases are less available than NIR databases [14], which impedes researches in the heterogeneous LWIR-VIS face recognition field.

Fig. 1. Different images of a same subject in different imagery bands (Visible and Infrared), from UL-FMTV Database [15].

We have chosen, in this paper, to take up all the challenges cited above. For that, we propose an effective and robust thermal LWIR/Visible heterogeneous face recognition approach that includes four main contributions:

- Full manual annotations of an existent thermal face database are proposed for the scientific community to develop future thermal face detectors.
- A Multi-sensors detector based on the recent YOLO v3 architecture is proposed for face detection in visible as well LWIR imagery.
- A Cycle GAN with modified loss function (called TV-CycleGAN) is proposed to synthesize visible faces from LWIR faces, in different real-world scenarios, reducing the cross-spectral modality gap.
- A promising face landmark detection in thermal images is proposed. It is based on those obtained from the generated face images using our proposed TV-CycleGAN.

II. Background

Several face recognition methods using two spectra have been developed over the last decade [7], [16], [17]. We can distinguish two main approaches: multispectral face recognition (MFR) and heterogeneous face recognition (HFR).

In the MFR approach, the facial recognition system is mainly based on a visible-infrared fusion step, which can be performed at three different levels: data level, feature level or match score level. The first level methods aim to obtain, from two or more coregistered images, one richer image that is used in the feature extraction step [18]–[20]. In the feature level fusion scheme, extracted characteristics from both spectra are merged in order to gather more informative and discriminative feature vectors [21]. This approach has been adopted to fuse extracted characteristics from both visible and LWIR or NIR images to construct a face recognition system in [22], [23]. For the score fusion scheme, after a score normalization process, classification scores are combined in order to improve the recognition performance [24].

In many real-world scenarios, we have images from the infrared spectrum only, since surveillance cameras often capture faces in low light conditions or in total darkness. However, most datasets accessible to law enforcement have been collected in the visible spectrum. Therefore, there exists a need to match IR images to visible face images. We aim in this work to deal with this heterogeneous face recognition challenge.

There are three main approaches in HFR field, which are common subspace, invariant features, and image synthesis. The first approaches aim to project heterogeneous domain spaces to a common subspace to ensure a better measure and more appropriate comparison than the original distributions. For example, the authors in [25] proposed a method, called Common Discriminant Feature Extraction (CDFE), where face images from near-infrared and visible sensors are projected into a common feature space, such that the intra-class gap is minimized and the inter-class gap is maximized. In [26] the canonical correlation analysis (CCA) is used in order to maximize the correlation between the near-infrared and visible imagery domains. Mapping all thermal LWIR and visible images onto a common subspace suffers from being computationally expensive and requires a big amount of pre-processing [27].

In invariant feature domain approaches, scientific researchers use local handcrafted features in order to compare face images gathered from different spectra. [28] for example, proposed light source invariant features (LSIF) to cancel heterogeneities for a NIR-Visible matching. Huang et al. [29] proposed three different modality invariant features, quantized distance vector (QDV), sparse coefficient (SC) and least square coefficient (LSC) as encoding to resolve a NIR-Visible heterogeneous face recognition problem. In the third approach, synthesized images in the reference domain are generated from a probe domain and vice versa to apply traditional homogeneous methodologies. In [27], a joint dictionary learning mapping has been proposed for image reconstruction from visible to NIR domain and vice versa. In [30], using a deep CNN network, visible faces are reconstructed from near-infrared images, by using a cross-modal hallucination at the input and a low-rank embedding at the output. In [31], a cascaded refinement network (CRN) has been proposed for LWIR to visible-like images synthesis. Kantarci et al [32] proposed to first apply the difference of Gaussians filter (DoG) and face alignment, using manually annotated facial landmarks, as a preprocessing stage, then they used a deep auto-encoder architecture based on U-Net network to learn the mapping from thermal face images to visible-like face images.

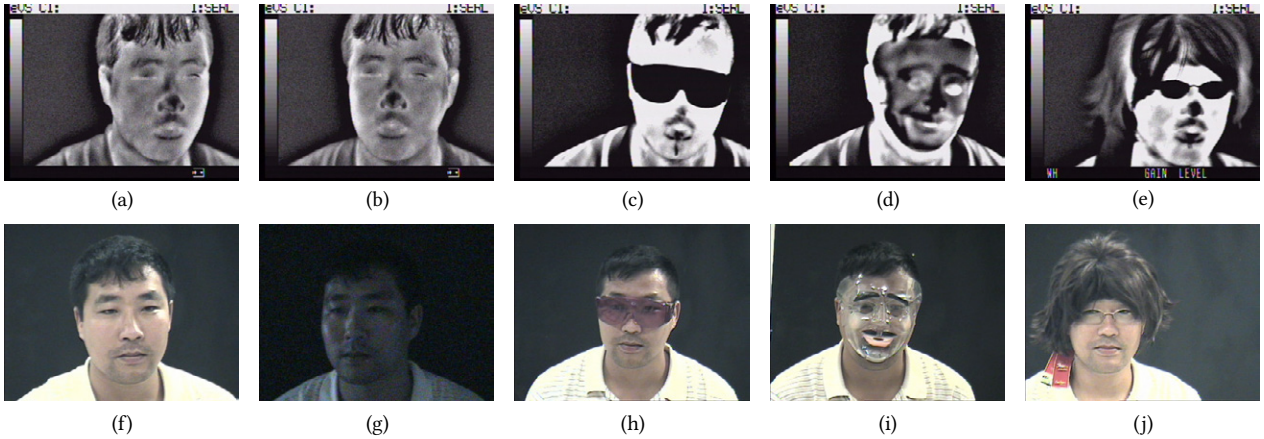


Fig. 2. Thermal imagery advantages under night and spoofing with disguise scenarios, (Up) LWIR Thermal (Down) Visible , (a,f) normal - (b,g) dark - (c,h) disguise with goggle - (d,i) disguise with mask - (e,j) disguise with wig.

Synthesis image approaches in HFR are based on cross-domain translation to reduce the modality gap. Therefore, we believe that generative adversarial network (GAN) [33] is an emerging and prosperous technique that may be applied in this field. Song et al [34], have introduced adversarial learning in raw-pixel and compact feature space to perform a NIR-VIS verification. In regard to the LWIR-VIS matching case, which could be considered as the most complicated case as aforementioned (lowest SSIM), there are very few scientific works in the literature. We can cite for example [35], where they have proposed the semantic guided GAN (SG-GAN) by assigning semantic labels, gathered from a face parsing network to semantic losses, to regularize the adversarial training. In [36], Zhang et al have proposed a TVGAN to generate realistic visible faces from LWIR thermal faces; however, all extremes poses and expressions were excluded from the experiences. Also, they did not provide any details related to automatic facial detection, which it is not a trivial task in LWIR imagery. Recently, Chu et al [37] proposed the multi-scale image synthesis method to translate thermal face images to the visible sensor modality. Their method is based on a GAN model to which they added the feature embedding, the facial landmarks and the identity preservation losses to their baseline loss function.

Other works are based on the polarimetric thermal infrared acquisition. This technique is used to achieve an improved performance since it retrieves the geometric and the textural face details. Xing et al [38] proposed the multi-scale attribute preserving GAN (Multi-AP-GAN) to synthesize visible face images from their corresponding polarimetric thermal face images for cross-modality face verification, by guiding the generator network with extracted attributes from the visible face images using a pre-trained VGG-Face network. He et al [39] proposed a visible face synthesis GAN-based (GAN-VFS) to generate visible faces from polarimetric thermal faces.

The Table I summarizes the related works of the synthesisbased image approaches in LWIR/Polarimetric to Visible HFR with their reported performances, including some comments and drawbacks.

In this scientific research, we propose a new thermal LWIR-VIS image synthesis method based on Cycle GAN, with a modified loss function. It includes all extreme poses and facial expressions in the evaluation protocol. In addition, as the cross-spectral translation depends highly on the pre-processing stage, we propose a simultaneous face detection technique in both spectra (visible and thermal LWIR) to automate the face-cropping phase. It is based on the last YOLO v3 deep architecture, trained using the WIDER visible face DB and the Terravic thermal DB, that we have annotated.

III. Proposed Approach

The proposed heterogeneous face recognition system is shown in Fig. 3. It includes three main parts:

- Multisensor Face Detection
- Face synthesis using TV-CycleGAN
- Face recognition

In addition to the database annotations, our scientific contribution concerns the first two parts.

We notice that the proposed TV-CycleGAN method used to generate visible faces from thermal images has a direct impact on face recognition process. We have chosen to focus on the thermal to visible face synthesis since most existing stored databases only contain visible face imagery of individual of interests [14]. In addition, the LWIR-VIS heterogeneous face recognition scenario is the most interesting for many security applications.

The results of the proposed face synthesis method using TVCycle GAN show also an interesting contribution in the face landmark detection field (Fig. 3). We will show more results in a later section.

A. The Proposed YOLO v3 Based Multi-sensors Face Detector

Face detection aims to locate the face coordinates in an image. Several face detection techniques have been proposed in the literature [41]. They can be categorized into two approaches: a region proposal approach and a regression/classification approach. The first approach, including RCNN [42], Fast-RCNN [43] and Faster- RCNN [44], adopts a pipeline of two stages: Object selection and classification. The first stage consists of selecting similar regions (same color, texture, or features). It is followed by a classification of each selected region. The regression/classification based approach, to which belong YOLO [45] and SSD [46] methods, is one-step straightforward. In this approach, images are firstly divided into grid cells where an object classification is carried out, reducing the time processing for region selection required in the first approach.

Unlike the R-CNN network versions that have used the regional proposal approach for the region selection, the YOLO network is faster and illegible for real-time applications, with comparable detection accuracy. YOLO v3 [47] is the last version, it incorporates the darknet 53 architecture, with skipped connections for feature extraction, and 53 other added convolution layers, for object detection, giving a fully convolutional network of 106 layers. Predictions are made at three different scales. As a result, this new architecture is more robust for the smallest objects than the early YOLO versions and has a better accuracy at overall.

TABLE I. Overview of the Related Works in Heterogeneous Face Recognition (Thermal to Visible) Using the Synthesis-Image Approach (RA: Recognition Accuracy, EER : Equal Error Rate)

Infrared to Visible	Reference (Year)	Synthesis Method	Pretrained Recognition Model	Database	Reported Performance	Comments/ Drawbacks
LWIR to Visible	[36] (2018)	TV-GAN	Mat-Conv-Net (VGG Face)	IRIS	RA: 13.9% (with 1 image/subject in the gallery) RA: 19.9% (with 4 images/subject in the gallery)	• Repeated angles, extreme poses, expressions, and illumination have been excluded from the experiments. - It produces low performances. • The equal Error Rate (EER) is not calculated.
	[31] (2019)	CRN	OpenFace	VIS-TH Eurecom	RA: 15.37%	• The explored database includes just 50 subjects. • The authors used Cascaded Refinement Networks (CRN) to consider multiple scales of images. It is different to generative adversarial networks concept. • Equal Error Rate (EER) is not calculated.
			LightCNN		RA: 57.612%	
	[35] (2019)	SG-GAN	AM-Softmax	Army Research Laboratory (ARL)	EER (LWIR to Visible): 14.24%	• As the number of subjects in the ARL dataset is not large, the recognition model has been trained on a larger MWIR dataset (PCSO), and fine-tuned using the ARL training set.
	[32] (2020)	DoG Filter + Autoencoder	Not Specified	Carl	RA: 48% (with 1 image/subject in the gallery) RA: 88.33% (with all images/subject in the gallery)	• The method uses autoencoders to learn Thermal-Visible mapping. • The used autoencoder is based on a modified U-Net architecture • The preprocessing stage does not include an automatic face detection stage. • The Carl database has been aligned using a manually annotated landmarks. • Equal Error Rate (EER) is not calculated.
				VIS-TH Eurecom	RA: 57.91% (with 1 image/subject in the gallery) RA: 85% (with all images/subject in the gallery)	
				UND-X1	RA: 58.75% (with 1 image/subject in the gallery) RA: 87.2% (with all images/subject in the gallery)	
	[37] (2021)	Multi-scale Image Synthesis	LightCNN	VIS-TH Eurecom	RA: 58.27%	Three losses have been added to the GAN loss function: feature embedding, identity preservation and facial landmarks losses • The equal error Rate (EER) is not calculated.
	[38] (2021)	Multi-AP-GAN	VGG-Face	VIS-TH Eurecom	EER: 25.68%	• The loss function of the generator network includes five losses: multi-scale adversarial loss, perceptual loss, identity loss, attributes loss and target-reconstruction loss.
				TUFTS	EER: 31.14%	
Polarimetric to Visible	[39] (2017)	GAN-VFS	VGG-Face	Army Research Laboratory (ARL)	EER (LWIR to Visible): 27.34 % EER (Polar to Visible): 25.17 %	• Polarimetric thermal imagers are very expensive for a daily life application compared to FLIR thermal imagers. • The polarimetric databases include geometric and textural details of faces that are not present in the thermal faces images.
	[40] (2018)	Multiple-Region Based Method	VGG-Face	Army Research Laboratory (ARL)	EER (LWIR to Visible): 26.25 % EER (Polar to Visible): 21.46%	
	[38] (2021)	Multi-AP-GAN	VGG-Face	Extended Army Research Laboratory (Ext-ARL)	EER (LWIR to Visible): 19.25 % EER (Polar to Visible): 17.81 %	

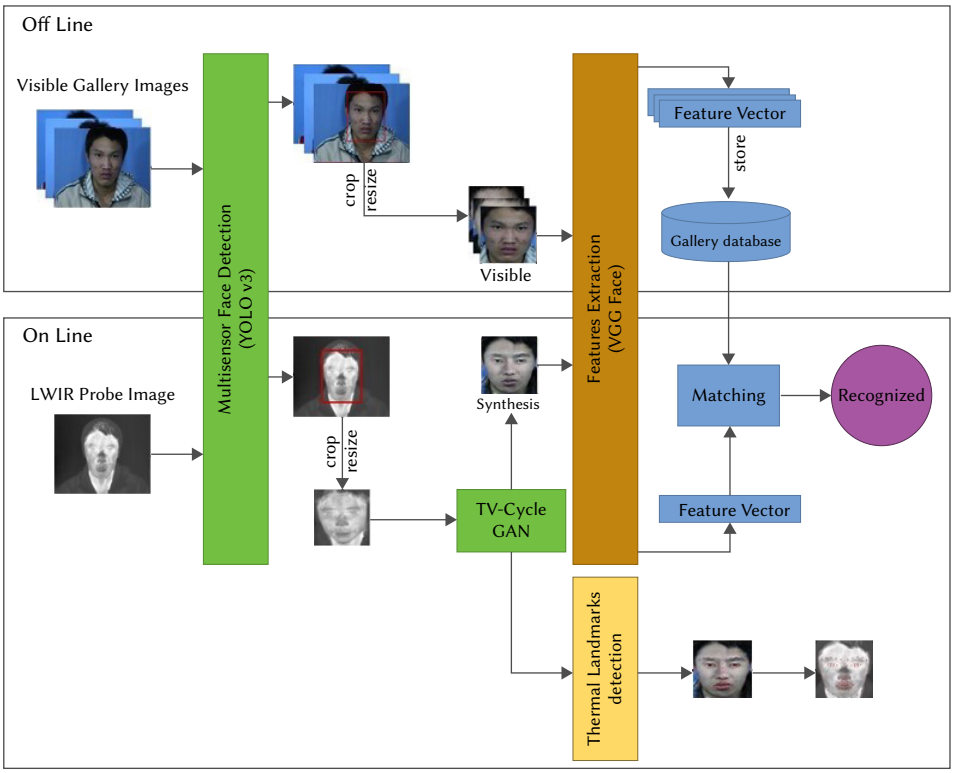


Fig. 3. Flowchart of the proposed Heterogeneous Face Recognition System.

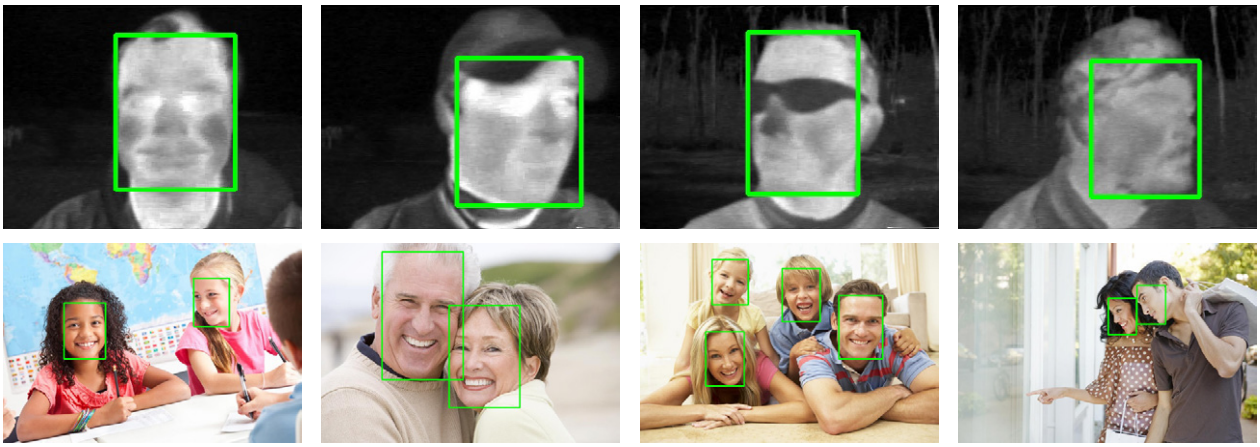


Fig. 4. Training labeled data Samples : (Top) Our manual face annotations regarding thermal LWIR imagery (Terravic DB), (Bottom) Visible face annotations (WIDER DB).

Face detection from visible images has been widely investigated in many scientific researches [48]–[50]. However, only a few works have been dedicated to thermal imagery [51], [52]. According to the best of our knowledge, there is no annotated public thermal face database that allows training a deep network for face detection. To overcome this issue, we have manually annotated the full thermal Terravic Facial IR database in the PASCAL VOC format (see the first row of Fig. 4) which contains a total of 21676 images.

In this paper, a multi-sensor face detection technique, based on the YOLO v3 network architecture, is proposed. It aims to detect faces in both visible and thermal images. For that, we have trained the network in two steps. Firstly, the Terravic Facial IR Database, which we have annotated, is used to train it for face detection in thermal images. Secondly, using a transfer learning technique and the WIDER database [53], a second full training phase has been carried out, for face detection in visible images. This database is fully annotated, including RGB face images with a high degree of variability in scale, pose, occlusion, expression, appearance, and illumination (Fig. 4).

B. Face Synthesis Using Thermal-Visible-Cycle Generative Adversarial Network (TV-CycleGAN)

Recently, Generative Adversarial Networks (GAN) have been proposed as an emerging area in deep learning field, based on two neural networks, a discriminator and a generator respectively. The discriminator acts as a binary classifier that discerns between real and fake images while the generator network learns epoch by epoch to produce images with realistic visual aspect as objective to fool the discriminator. According to application needs, variant GAN architectures have been proposed in the literature such as cGAN [54], Pix2Pix [55] or Cycle GAN [56]. In this work, we are interested in Cycle GAN.

Cycle GAN is based on a pair of vanilla GANs including in total two generators and two discriminators, denoted by (G_T; G_V) and (D_T; D_V) respectively. In order to transform thermal face images t to like visible face images $\hat{v}$, the generator G_T learns the domain translation T (Thermal) to V (Visible) based on the adversarial loss described in equation (1). The negative log-likelihood objective is replaced by a least-squares adversarial loss [57] in order to avoid gradient vanishing problem for training stability and quality generation as mentioned in the original Cycle GAN paper [56], giving the objective loss function described in equation (2). The opposite GAN direction, from domain V (Visible) to T (Thermal), is learned by (G_V ; D_T) according to equation (3).

$$\mathcal{L}_{cyc}(G_V, G_T) = \mathbb{E}_{t\sim p_{data}(t)}\|G_V(G_T(t)) - t\| + \mathbb{E}_{v\sim p_{data}(v)}\|G_T(G_V(v)) - v\| \quad (1)$$

$$\mathcal{L}_{GAN(T\to V)}(G_T, D_V, T, V) = \mathbb{E}_{v\sim p_{data}(v)}[(D_V(v) - 1)^2] + \mathbb{E}_{t\sim p_{data}(t)}[D_V(G_T(t))^2] \quad (2)$$

$$\mathcal{L}_{GAN(V\to T)}(G_V, D_T, V, T) = \mathbb{E}_{t\sim p_{data}(t)}[(D_T(t) - 1)^2] + \mathbb{E}_{v\sim p_{data}(v)}[D_T(G_V(v))^2] \quad (3)$$

The main differences between Cycle GAN and Classical GAN are cycle consistency loss Lcyc and identity loss Lid described in equations (4) and (5) respectively. Similar to conventional autoencoders, L_{cyc} corresponds to the L_1 norm of image reconstruction. In case of $T \to V$ transformation, by injecting a newly generated visible like image $\hat{v}$ from thermal image t using G_T, through the second generator network G_V, leads to reconstruct the original image t, ie: $t \to G_T(t) \to G_V(G_T(t))$ t. In order to reduce the translation domain space, the cycle consistency loss is added to the GAN loss function.

The identity loss L_{id} refers to the L_1 norm between the input image and the generated image mapped from its domain, added to preserve the color composition and identity features. The final Cycle GAN loss function is described in equation (6). λ and α are two fixed parameters to control the loss impact on the objective function.

$$\mathcal{L}_{cyc}(G_V, G_T) = \mathbb{E}_{t\sim p_{data}(t)}\|G_V(G_T(t)) - t\| + \mathbb{E}_{v\sim p_{data}(v)}\|G_T(G_V(v)) - v\| \quad (4)$$

$$\mathcal{L}_{id}(G_V, G_T) = \mathbb{E}_{t\sim p_{data}(t)}\|G_T(t) - t\| + \mathbb{E}_{v\sim p_{data}(v)}\|G_V(v) - v\| \quad (5)$$

$$\mathcal{L}_{CycleGAN}(G_V, G_T, D_V, D_T) = \mathcal{L}_{GAN(T\to V)} + \mathcal{L}_{GAN(V\to T)} + \lambda\mathcal{L}_{cyc} + \alpha\mathcal{L}_{id} \quad (6)$$

Referring to [14], which the structural similarity index metric (SSIM), defined in formula (7), is used to study the modality gap between visible and infrared sub-bands, we have opted to add the reverse metric (equation 8) to the CycleGAN's loss function to reduce this gap and improve the visual aspect of domain translation from thermal infrared to visible, giving the TV-CycleGAN loss, described in equation (9).

$$SSIM(v, t) = \frac{(2\mu_v\,\mu_t + b_1)(2\sigma_{vt} + b2)}{(\mu_v^2 + \mu_t^2 + b_1)(\sigma_v^2 + \sigma_t^2 + b_2)} \quad (7)$$

$$\mathcal{L}_{SSIM(G_V,G_T,v,t)} = (1 - SSIM(G_T(t), v)) + (1 - SSIM(G_V(v), t)) \quad (8)$$

$$\mathcal{L}_{TV-CycleGAN}(G_V, G_T, D_V, D_T) = \mathcal{L}_{CycleGAN}(G_V, G_T, D_V, D_T) + \mathcal{L}_{SSIM} \quad (9)$$

where (μ_v; μt) and (σ_v^2; σ_t^2) denote the mean and the variance, of the respective images v and t. σ_{vt} refers to the cross-covariance. b_1 and b_2 are two constants added to avoid instability when ($\mu_v^2 + \mu_t^2$) or ($\sigma_v^2 + \sigma_t^2$) are close to zero.

The proposed TV-CycleGAN pipeline's details are given in Fig. 5.

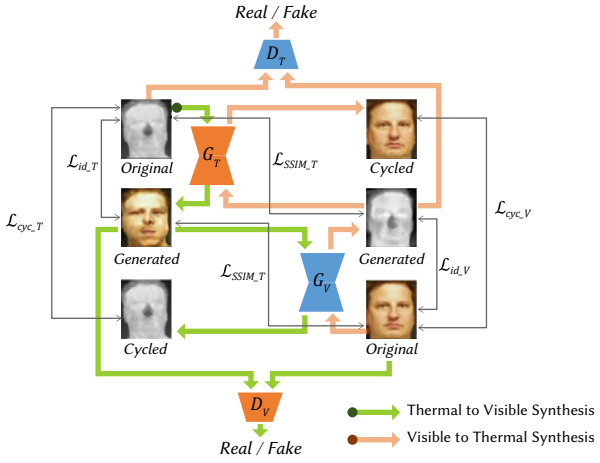


Fig. 5. Flowchart of the proposed TV-CycleGAN, where $L_{cyc} = L_{(cyc_V)} + L_{(cyc_T)}$, $L_{id} = L_{(id_V)} + L_{(id_T)}$ and $L_{SSIM} = L_{(SSMI_V)} + L_{(SSMI_T)}$.

C. Face Recognition

The proposed TV-CycleGAN is a technique to translate LWIR face images to visible like face images and vice versa, based on a modified adversarial loss. However, the identity information must be preserved to recognize a subject by visible-visible or LWIR-LWIR matching.

For the face recognition task (Fig. 3), we have used two pretrained models on a large-scale face dataset as a feature extraction technique; VGG 16 and Resnet 50. First, the extracted features from the real visible images of the TV-CycleGAN's testing subset, are used to construct the reference embedding. Afterwards, other testing embeddings with the same technique, are formed based on the extracted features from the synthesized images (these synthesis images are obtained from LWIR images using Pix2Pix, TV-GAN, CycleGAN and TV-CycleGAN, respectively).

To evaluate the identity preservation, the nearest neighbour is applied for a fair comparison with the state-of-the-art methods [31], [35], [36], [38]. It is the most commonly used matching process.

It uses the cosine distance, which is calculated for each feature vector in the testing embeddings with those from the reference embedding. Each feature vector from the test set is then classified with the label referred to the lowest distance in the reference embedding and the recognition accuracy is calculated.

The details of each used model are given in the following, and illustrated in the Fig. 6.

Fig. 6. Details regarding each used pretrained model for the face recognition stage : (Top) VGG16, (Bottom) RESNET 50.

1. VGG16

The VGG16 architecture [58] encompasses 16 deep layers, distributed through five blocks and followed each by a maxpooling layer with a size of 2×2 . The first block comprises two conv layers of 64 filters, and similarly for the second with 128 filters. The following blocks include 3 conv layers each, with 256 filters for the third one and 512 filters for the fourth and fifth blocks. The conv filters use a kernel size of 3×3 with the rectified linear unit (ReLu) activation function. The blocks end with three fully connected layers and a Softmax for prediction probabilities.

For our experiments, the VGG16 model was trained on the VGG Face dataset [59], which includes 2.6 million face images from 2622 subjects. We kept the model up to the fifth block, yielding to extract feature vectors of size 512 attributes.

2. RESNET50

The RESNET 50 belongs to the residual networks [60], including shortcut connections. The network starts with a block of 64 conv filters with kernel size of 7×7 and a maxplooing layer of size of 3×3 with a stride of 2. The rest of the network is distributed through four other blocks of 3 conv layers inter-connected using a shortcut connection each, and kernel sizes of 1, 3 and 1, respectively. The second and fifth blocks are repeated three times while the third, 4 times, and the fourth 6 times, giving in total 50 layers. It is followed by an average pooling layer and fully connected layers of 1000 nodes with a Softmax.

For our experiments, the RESNET 50 model was trained on the VGG Face 2 dataset [61] that comprises 3.31 million face images gathered from 9131 identities. We erased to last layers to extract feature vectors by the last block with 2048 attributes.

IV. Experimental Results

A. Used Material and Tools

All our networks have been implemented using the Keras Python framework. The proposed face detector deep network is trained using the following configuration: learning rate of 0.0001 with Adam optimizer, callbacks such as early stopping and learning rates decay by factors 0.1 for 5 and 2 consecutive unimproved epochs, respectively. A batch size of 4 has been fixed to prevent the used graphical card, which is Nvidia GeForce RTX 2080 GPU with 8GB GDDR6, from overloading.

All our networks have been implemented using the Keras Python framework. The proposed face detector deep network is trained using the following configuration: learning rate of 0.0001 with Adam optimizer, callbacks such as early stopping and learning rates decay by factors 0.1 for 5 and 2 consecutive unimproved epochs, respectively. A batch size of 4 has been fixed to prevent the used graphical card, which is Nvidia GeForce RTX 2080 GPU with 8GB GDDR6, from overloading.

Our TV-CycleGAN adopts the U-Net [62] architecture for its generator, with skipped connections. We have adapted the network to our input images of resolution 128×128 px and used blocks of convolution filters, leaky relu as activation function and instance normalization. For the discriminator network, a fully convolutional neural network using the same blocks as the generator has been used for the binary classification. The TVCycleGAN was trained from scratch for 200 epochs, with the default parameters of the original CycleGAN paper: $\lambda = 10$ and $\alpha = 1$, learning rate of 0.0002 with Adam optimizer and a batch size of 1. Nvidia Tesla K80 was used in this experiment.

B. Used Datasets

Five databases have been used in the experiments that we have carried out; they are presented in the following.

1. Terravic Thermal Database

Terravic Thermal DB[1] provides a set of thermal face images, for 20 different individuals, captured using the Raytheon -3 Thermal-Eye 2000AS thermal sensor, under different scenarios (normal posture, wearing sunglasses/hats, indoor/outdoor and variations in pose). It consists of a total of 21676 images with resolution of 320×240 px. The face images of each subject are put in one folder and split into two subsets: a training set (face folders from 10 to 20), including 16462 images, and a test set of 5197 images (face folders from 01 to 04, and 07 to 09), folders 05 and 06 are not available.

2. WIDER Database

WIDER DB [53] is a public fully annotated database in PASCAL VOC format, it consists of 32203 visible images from 61 different event classes. These images contain a total of 393703 labeled faces including occlusions and several variations in scale and pose. The database is divided into 3 subsets, 40% for training (12880 images), 10% for validation (3226 images) and the remaining 50% for the test set.

3. TUFTS Database

TUFTS DB [63] is one of the largest public[2] databases that provide faces images acquired in different modalities such as visible, thermal, near-infrared, 3D and facial sketch. It consists of over 10000 images, collected from 112 individuals from more than 15 countries, with several age groups. These images include facial expressions and various acquisition angles. In our experiments, we have used the visible and thermal subsets, which include in total 1537 paired images. The thermal images in this database were acquired using the FLIR Vue Pro camera. A new version of this database has been shared in 2020. It includes the same pair images (LWIR-Visible) as the original TUFTS database but in aligned form. In this paper, we have called this database "Aligned TUFTS".

4. NVIE Database

USTC-NVIE[3] is a multispectral database that comprises spontaneous and posed facial expressions [64], [65]. The posed sub-database consists of over than 7000 face images of 107 different subjects, gathered from the visual and LWIR spectra simultaneously at a distance of 0.75m, using the DZGX25M visible camera and the SAT-HY6850 infrared camera respectively. The images include variations in illuminations and facial expressions.

5. PUJ Database

Pontificia Universidad Javeriana (PUJ) [66] database is a public multispectral face dataset that provides 800 paired and nonaligned images from both visible and LWIR modalities. It was acquired from 40 subjects using the FLIR-T360 according to two illumination protocols (with/out lighting), including variations in pose (front and profile capture) and facial expressions (neutral, surprised, and smiling).

The overview of each used database is given in Table II.

TABLE II. Overview of the Used Face Databases

Database	# Subjects	# Images	VIS	NIR	LWIR
Terravic DB	20	21676	×	×	✓
WIDER DB	-	32203	✓	×	×
TUFTS DB	113	+10000	✓	✓	✓
NVIE DB	107	7329	✓	×	✓
PUJ-T360 DB	40	800	✓	×	✓

[1] http://vcipl-okstate.org/pbvs/bench/Data/04/download.html

[2] http://tdface.ece.tufts.edu/

[3] https://nvie.ustc.edu.cn/

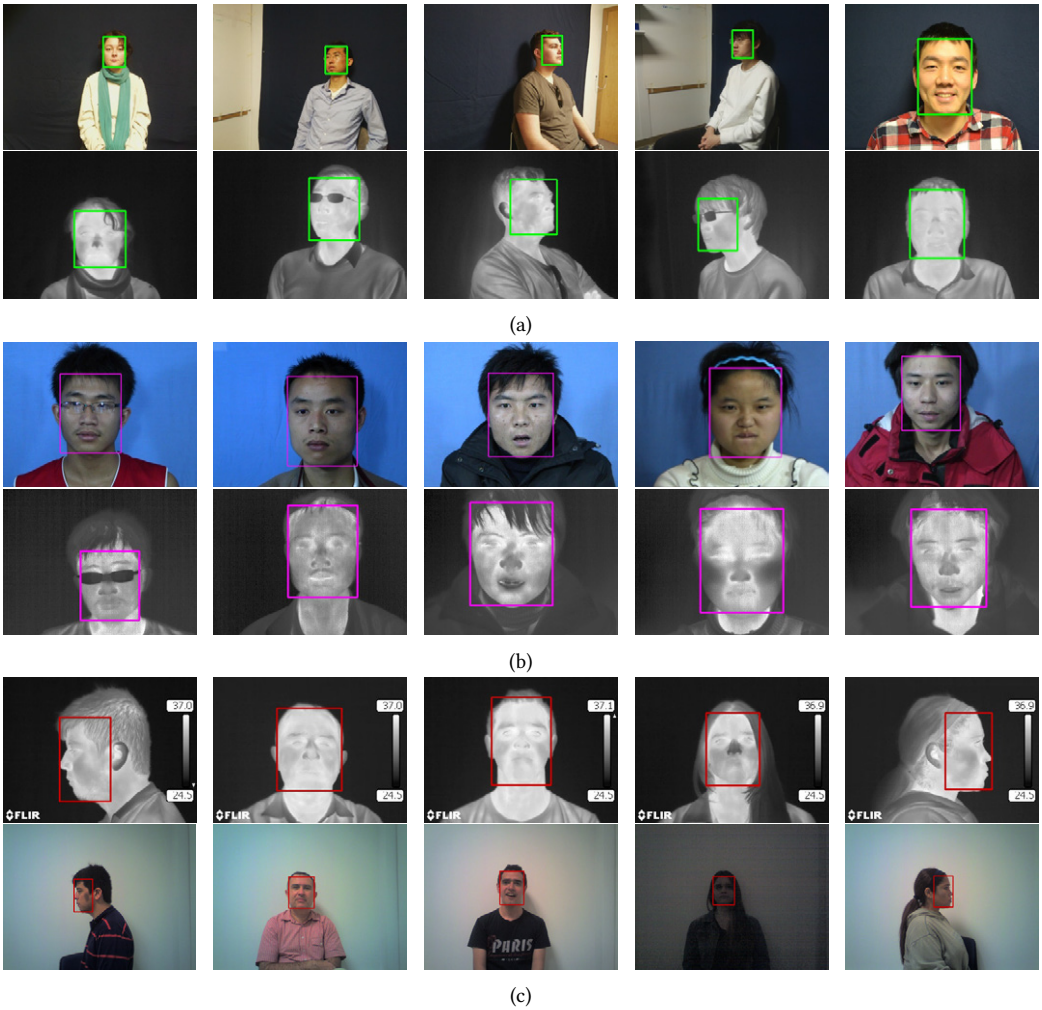


Fig. 7. Results of the proposed multi-sensor face detection technique, applied in different environment scenarios, such as normal posture, presence of facial expressions, slight and extreme angle orientations, and illumination changes, to visible and their corresponding thermal face images, (a) TUFTS database, (b) NVIE database, (c) PUJ database.

C. Protocols

As mentioned beforehand, our new proposed heterogeneous face recognition approach includes three parts: automatic multispectral face detection, visible synthesis using TV-CycleGAN and face recognition (Fig. 3).

Firstly, in the multi-spectral detection step, we have trained a custom YOLO v3 for face detection, using our Terravic face annotations for the thermal modality and the WIDER face annotations for the visible modality. The training process has been carried out in two stages: first, a thermal face detector has been trained using only our Terravic annotations; then, by applying a transfer learning technique, the learned weights have been trained again using the WIDER's face annotations. The proposed multi-spectral face detection technique has been used in the prepossessing stage of our TV-Cycle GAN to automatically crop the region of interest (ROI) for the thermal to visible face translation.

Secondly, in the LWIR to visible translation step using our TVCycle GAN, three multi-spectral databases have been used for evaluation: TUFTS, NVIE and PUJ databases. All thermal/visible face images pairs have been automatically cropped using our own YOLO v3 model and then have been split randomly into two subsets (training/test), according to ratios equal to 95/17 and 92/15, for the TUFTS and NVIE databases, respectively.

Concerning the PUJ database. As it is the smallest one, we have considered the TUFTS database as the training set and we have used all the PUJ images for the test (as probe images). This procedure allows us to test the robustness and the efficiency of the proposed method when the system and acquisition conditions change, since the FLIR Vue Pro has been used for the TUFTS database acquisition where the FLIR T360 has been used for the PUJ database acquisition. Also, we have carried out experiments on Aligned TUFTS database because it is one of the newest available datasets with aligned images, which constitutes an interesting way to show the recognition rate enhancement introduced by our method for aligned faces.

Finally, in the face recognition step, the real-visible face images are enrolled into a pretrained VGG face network, based on the VGG 16 [59] and the RESNET 50 architectures [61], to construct the reference face embedding. Afterwards, in the testing phase, the extracted feature vectors from each probe synthesized-visible image, are compared against those stored in the reference embedding for a face identification related the lowest cosine distance.

Furthermore, we have applied facial landmarks detection on the obtained synthesized images using the proposed TVCycleGAN method. From that, we have easily detected the landmarks on the original thermal ones. Thanks to this result, we have dealt with thermal face landmark detection which is one of the most challenging task.

TABLE III. Quantitative Evaluation of the LWIR to Visible Face Images Synthesis Regarding TUFTS and NVIE Databases

Method	TUFTS					NVIE				
	SSIM	PSNR	MSE	RMSE	MAE	SSIM	PSNR	MSE	RMSE	MAE
Raw Thermal	0.305	9.932	7016.4	82.48	66.7	0.221	10.195	6400.9	79.42	67.28
Pix2Pix	0.332	12.906	3569.6	58.74	44.25	0.274	13.635	2977.6	53.81	42.74
TV-GAN	0.321	12.761	3671.4	59.61	45.22	0.271	13.031	3434.7	57.74	46.34
CycleGAN	0.381	13.902	2978.4	53.03	40.41	0.307	14.254	2510.7	49.75	38.44
TV-CycleGAN (Our)	**0.384**	**13.964**	**2902.3**	**52.50**	**39.97**	**0.324**	**14.28**	**2473.8**	**49.36**	**37.9**

D. Performance Evaluation

1. Face Detection Results

The face detection results obtained by our YOLO v3 based network are illustrated in Fig. 7, using different visible face images and their corresponding LWIR face images. Several acquisition scenarios have been considered to evaluate our multisensors face detector: front pose, slight and extreme orientation, high and low illumination conditions. These results clearly demonstrate the efficacy of the proposed face detector, under the variations cited above.

2. Visible Synthesis Results

To show the performance of our TV-CycleGAN proposed method, for visible face image synthesis, several experiments have been carried out using three multi-spectral face databases, called TUFTS and its aligned version, NVIE and PUJ, respectively. Also, quantitative, qualitative evaluations and comparison to other state-of-the-art methods have been done.

The quantitative evaluation has consisted in computing the Structural Similarity Index (SSIM), the Peak Signal to Noise Ratio (PSNR), the Mean Squared Error (MSE), the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) metrics, for each pair of thermal-visible images and each pair of synthesizedreal visible images, for the TUFTS and NVIE databases. The average value of each metric is reported in Table III.

We notice from the thermal-visible SSIM related to the used databases, that we deal in this paper with the most complicated case, because the LWIR-VIS modality gap corresponds to an SSIM of 0.305 for TUFTS DB and 0.221 for NVIE DB. These gaps are higher than the one reported in the comparative study of [14] where the SSIM is equal to 0.335.

The comparison of different GAN-based methods: Pix2Pix [55], TV-GAN [36], Cycle-GAN [56], and TV-CycleGAN; using the metrics SSIM, PSNR, MSE, RMSE and MAE, shows that the proposed method (TV-CycleGAN) allows the best modality gap reduction for both databases (Table III).

Even if the difference seems small between the results of CycleGAN and TV-CycleGAN (all metrics at Table III), the TVCycleGAN provides a significant improvement on the quality of synthesized faces as shown in Fig. 8 and 9 related to the qualitative evaluation.

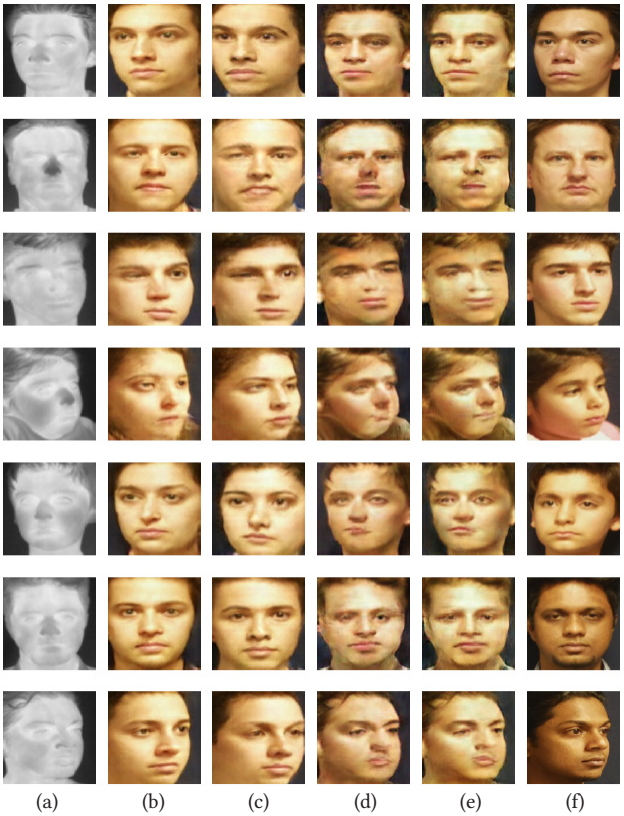


Fig. 8. Qualitative comparative study regarding TUFTS database for the visible face synthesis from LWIR face images: (a) Plain Thermal, (b) Pix2Pix [55], (c) TV-GAN [36], (d) CycleGAN [56], (e) TV-CycleGAN (Ours), (f) Target Visible.

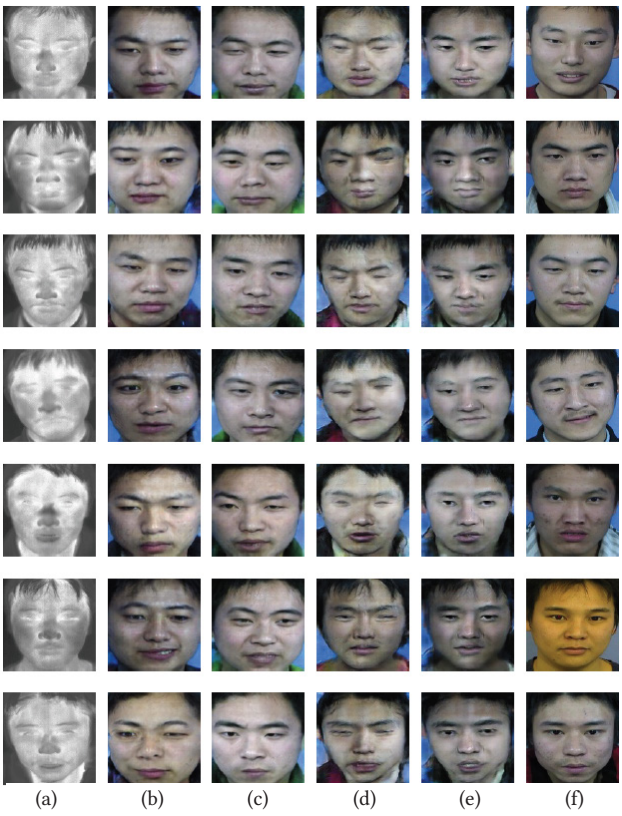


Fig. 9. Qualitative comparative study regarding NVIE database for the visible face synthesis from LWIR face images: (a) Plain Thermal, (b) Pix2Pix [55], (c) TV-GAN [36], (d) CycleGAN [56], (e) TV-CycleGAN (Ours), (f) Target Visible.

The qualitative evaluation is based on visual inspection. We have randomly selected some face image samples to show the results of our proposed visible synthesis method, as shown in Fig. 8 and 9. Even though all synthesizing methods fail sometimes to generate the detailed facial attributes, due to the absence of regularization guiding the GAN training, the proposed one, whose results are shown in the fifth column, outperforms the other state-of-the-art methods and provides a satisfying generation quality. Indeed, the synthesized images shown in this column preserve the persons' identities, they are close to their corresponding ground truth images, shown in column 6.

We can notice from the results related to TUFTS database (Fig. 8), that the TV-CycleGAN, provides a satisfying generation quality, thanks to the loss function similarity incorporated, as it has been mentioned beforehand. Even if the difference seems small between the results of CycleGAN and TV-CycleGAN (Table III), the TV-CycleGAN provides a significant improvement on the quality of synthesized faces. As shown in Fig. 8, 9 and 10 related to the qualitative evaluation, the proposed TV-CycleGAN gives more realistic faces and conserves better than other methods person's identity. For example, in row 2, the TV-CycleGAN is able to better generate nose attributes. In row 4, our method generates the face with better mouth and nose than Cycle-GAN. Also, in row 5, the obtained face from TV-CycleGAN has a better generated mouth.

Similarly, for the NVIE database (Fig. 9), our TV-CycleGAN outperforms the CycleGAN for facial attributes generation in all cases, particularly the eyes, eyebrows, nose, and skin texture, as shown in Fig. 10.

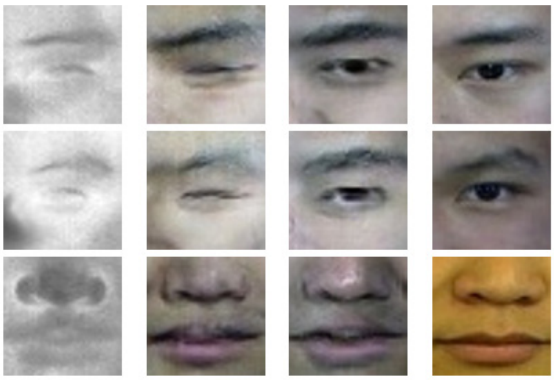


Fig. 10. Enlarged regions of facial attributes, eyes and eyebrows, nose and mouth from Fig. 7 to compare TV-CycleGAN against its main competitor CycleGAN: (a) Plain Thermal. (b) CycleGAN (c) TV-CycleGAN, (d) Target Visible.

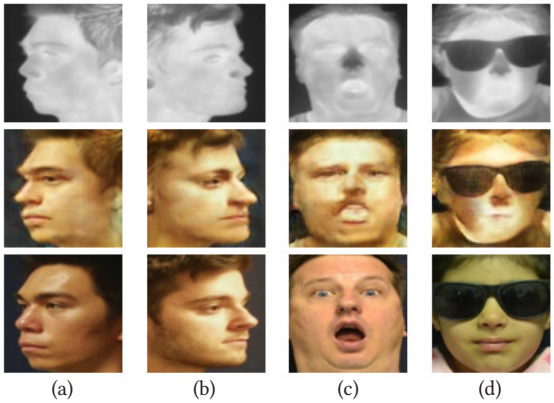


Fig. 11. LWIR to Visible translation using TV-CycleGAN for scenarios including: extreme poses ((a) and (b)), facial expression (c) and glasses (d). first row: Raw Thermal, second row: TV- CycleGAN transformation, third row: Target Visible.

To reinforce the efficacy of the proposed method, we have included extreme poses and facial expressions in our TVCycleGAN training phase (Fig. 11).

In the light of all these results, we can assume that the proposed TV-CycleGAN will have a great impact on the face recognition results as shown in the next section.

3. Face Recognition Results

The proposed TV-CycleGAN's main goal, is reducing the LWIR-Visible modality gap, by synthesizing visible-like face images from thermal ones while preserving as much as possible relevant identity information, to improve the accuracy of heterogeneous face recognition.

In the face recognition step, we have used the pretrained models; VGG16 [59] and the RESNET50 [61] for feature extraction. First, a reference face embedding has been built with the extracted features from the testing real-visible subset images. Afterwards, each synthesized image from LWIR using GANs, is enrolled into one of the pretrained models and then classified with the label of the closest feature vector in the reference embedding, based on the cosine distance. Four metrics have been used to evaluate the face recognition performances as shown in Tables IV and V.

From Table IV, one can figure out that the modality gap has a huge impact on the identification accuracy, since matching directly raw thermal face images with the visible ones, gives a low accuracy, which is 28.99 % in the case of the TUFTS DB and only 17.06 % in the case of the NVIE DB, corresponding to 23.48 % and 8.80 % respectively in terms of F1-Scores, when the model VGG 16 is used.

Synthesizing visible-like images to match the real-visible ones often improves the heterogeneous face recognition performance. The TV-CycleGAN achieved the best recognition rates, which are 57.56 % for the TUFTS DB using the VGG 16 model, and 58.32% for the NVIE DB using the RESNET 50 model. Compared to its main competitor, Cycle GAN, our proposed TV-CycleGAN allows an improvement rate of 0.42% and 1.52% for the TUFTS and NVIE databases respectively. Comparing to a direct LWIR-Visible face matching results, the proposed TV-CycleGAN allows an accuracy enhancement of 10.5 % for TUFTS database, and 29.16% for NVIE database.

This improvement results from the new loss function implemented in the TV-CycleGAN. These results are in good agreement with the quantitative evaluation reported in Table 3; they prove that the modality gap reduction brought by the TVCycleGAN contributes to improve the accuracy of heterogeneous face recognition systems (Tables IV and V). Indeed, compared to the Pix2Pix [55] and TV-GAN [36] methods, the proposed TVCycleGAN has the largest AUC and the lowest EER, for both databases, as shown in Tables 4. In addition, to visualize the contribution of TV-CycleGAN synthesis on the face recognition performances, we have applied the Grad-CAM [67] on the TUFTS database as shown in Fig. 12. This shows that the TV-CycleGAN allows the recognition model to focus more on the facial attributes with larger activated regions (referred to the red regions) compared to the original LWIR face images. The activated regions are closer to the ones activated on the ground truth faces, which demonstrates the accuracy improvements compared to the lowest results obtained with the LWIR to visible face matching.

Furthermore, the obtained results from Table V for the aligned TUFTS database show that an additional face alignment stage contributes to improve the recognition performances. Therefore, the TV-CyleGAN reaches the top accuracy of 63.45 %. This result corresponds to an accuracy enhancement of 18.49% in comparison to a direct Aligned LWIR-Visible face matching. On the other hand, our method outperforms the recent Multiple-APGAN one [38] with an enhancement of 3.92% in terms of EER, even if this last one uses more

TABLE IV. Obtained Face Recognition Results Regarding the TUFTS and NVIE Databases

Method	TUFTS				NVIE			
	Accuracy	F1-Score	AUC	EER	Accuracy	F1-Score	AUC	EER
Raw Thermal – VGG 16	28.99%	23.48%	62.28%	42.63%	17.06%	8.80%	55.57%	46.85%
Raw Thermal – RESNET 50	47.06%	42.99%	71.88%	35.38%	29.16%	27.23%	62.05%	42.73%
Pix2Pix – VGG 16	49.57%	49.69%	73.21%	34.25%	23.97%	16.88%	59.27%	44.56%
Pix2Pix – RESNET 50	39.50%	39.03%	67.86%	38.61%	17.28%	13.05%	55.68%	46.78%
TV-GAN – VGG 16	47.90%	46.15%	72.32%	35.00%	14.25%	9.26%	54.07%	47.74%
TV-GAN – RESNET 50	43.70%	43.46%	70.09%	36.85%	17.28%	14.08%	55.68%	46.78%
CycleGAN – VGG 16	57.14%	55.53%	77.23%	30.57%	46.00%	42.09%	71.07%	35.96%
CycleGAN – RESNET 50	52.94%	54.07%	75.00%	32.65%	56.80%	55.40%	76.86%	30.83%
TV-CycleGAN – VGG 16 (Our)	**57.56%**	55.17%	**77.46%**	**30.36%**	46.65%	45.00%	71.42%	35.68%
TV-CycleGAN – RESNET 50 (Our)	55.88%	**55.79%**	76.56%	31.21%	**58.32%**	**55.60%**	**77.67%**	**30.05%**

TABLE V. Obtained Face Recognition Results Regarding the Aligned TUFTS Version and PUJ Databases

Method	Aligned TUFTS				PUJ			
	Accuracy	F1-Score	AUC	EER	Accuracy	F1-Score	AUC	EER
Raw Thermal – VGG 16	29.41%	22.93%	62.50%	42.48%	14.05%	7.97%	55.92%	46.78%
Raw Thermal – RESNET 50	44.96%	39.43%	70.76%	36.31%	27.27%	24.08%	62.70%	42.56%
CycleGAN – VGG 16	58.40%	55.55%	77.90%	29.93%	25.62%	20.16%	61.86%	43.12%
CycleGAN – RESNET 50	62.60%	61.17%	80.13%	27.69%	42.15%	37.68%	70.33%	36.99%
TV-CycleGAN – VGG 16 (Our)	**63.45%**	**61.51%**	**80.58%**	**27.22%**	34.71%	27.81%	66.52%	39.90%
TV-CycleGAN – RESNET 50 (Our)	63.03%	59.42%	80.35%	27.46%	**42.98%**	**37.78%**	**70.75%**	**36.65%**

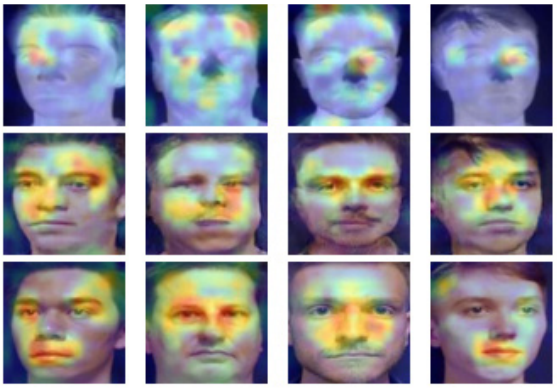


Fig. 12. Grad-CAM heatmaps from the TUFTS database using the pretrained VGG16 model. First row: Raw Thermal, second row: TV-CycleGAN transformation, third row: Ground truth.

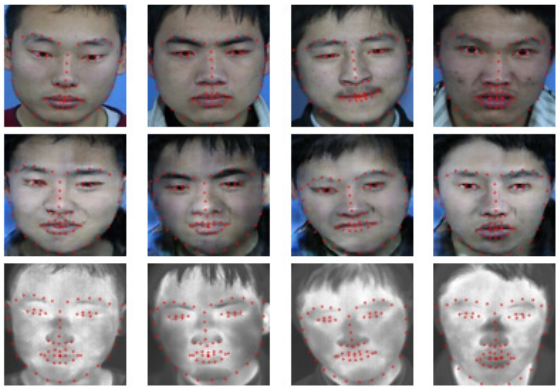


Fig. 13. Thermal facial landmark detection results based on TV- CycleGAN transformation. Top: detection on real visible images, Middle: detection on synthesized visible images from LWIR, Bottom: transferred facial landmarks coordinates on Thermal face images from those detected on synthesized face images (Middle).

losses (5 losses) in its objective function than our method that uses only one additional loss (Eq. 9).

Concerning PUJ database, the proposed TV-CycleGAN method shows its robustness and efficacy when the system and acquisition conditions change. Indeed, it reaches the accuracy rate of 42,98% and contributes to an accuracy enhancement of 15.71% compared to the LWIR to Visible face matching.

4. Thermal Facial Landmark Detection Results

We have also dealt with one of the hot topics in LWIR imagery, which is thermal face landmarks detection [68]–[70]. It is quite hard to directly perform the detection in thermal imagery due to its low contrast. To overcome this issue, our proposed TV-CycleGAN method proves its ability to allow interesting facial landmarks detection on the generated face images. These landmarks could be applied directly on the thermal ones. For that, we have applied the dlib Python library to detect 68 landmarks on the synthesized images, and then we have projected them on the LWIR face images, as shown in Fig. 13. As it can be observed, the obtained results are very satisfactory. There are very promising for many applications such as face tracking and automatic multispectral face alignment.

5. Complexity and Time Computation Results

To show the computation performance of our proposed heterogeneous face recognition, two hardware configurations have been used. The first configuration consists of Intel i3 5010 as CPU without any dedicated GPU where the second one uses the Intel Xeon CPU with Tesla K80 GPU. We have computed the complexity of each model used in each stage based on the number of floated point operations (FLOPs) and the corresponding average time elapsed for computation. The obtained results are reported in Table VI.

From Table VI, we can figure out that our heterogeneous face recognition system can be adapted easily to video surveillance systems including its three stages, and can run at a frame rate of 5 FPS when a dedicated GPU is used; however, when the hardware configuration has only a CPU for the computation, it takes 2.78 seconds to detect, synthesize and recognize the subject with the VGG16 model that corresponds to 83.65×10^9 FLOPs, and 2.49 seconds with the RESNET 50 model, corresponding to 72.15×10^9 FLOPs, respectively. This is unsuitable for real-time applications.

TABLE VI. Time Computation Performance of Our Heterogeneous Face Recognition System Regarding Two Hardware Configuration With/ Without GPU

Stages	Model	FLOPs	Configuration 1	Configuration 2
Face Detection	YOLO v3	65.86×10^9	1895.72 ± 147.51ms	153.08 ± 27.42ms
Face Synthesis	TV-CycleGAN	2.49×10^9	80.53 ± 5.7ms	7 ± 0.6ms
Face Recognition	VGG16	15.3×10^9	804.66 ± 151.01ms	32.03 ± 3ms
	RESNET50	3.8×10^9	508.94 ± 102.32ms	11.95 ± 0.8ms
Total	with VGG16	83.65×10^9	2780.91 ± 304.22ms	192.11 ± 31.02ms
	with RESNET50	72.15×10^9	2485.19 ± 255.53ms	172.03 ± 28.82ms

V. Conclusion

Cross spectral face recognition is a challenging task due to the large gap between modalities, especially between LWIR and visible spectra. Our contribution in this paper is fourfold. First, we have proposed a thermal-visible face detection method. This method is based on the YOLO v3 architecture and provides an advanced solution for face detection in both thermal and visible imagery, which makes it suitable for several applications, such as facial emotion recognition (FER) or liveness detection. Second, We have annotated a full thermal face database and have shared it with the scientific community in Github repository[4]. Third, we have proposed a modified CycleGAN, called, TVCycleGAN that allows to translate LWIR images to visible-like images. Finally, the synthesized-visible face images obtained by this network are very promising for thermal facial landmark detection. Summing up the results, it can be concluded that the proposed TV-CycleGAN deep learning network shows its robustness and efficacy in LWIR to visible faces synthesis for heterogeneous face recognition. Compared to some recent stateof- the-art methods, the proposed method gives more realistic faces and conserves better persons' identities. Thus, our results could be considered as a major asset for the use of our proposed system in daily real-life scenarios. For the future, we plan to improve the proposed method by guiding the GAN training. We also aim to extend our findings to Thermal-Visible face registration, using the detected face landmarks in synthesized visible images.

References

[1] N. K. Benamara, M. Keche, M. Wellington, Z. Munyaradzi, "Securing E-payment Systems by RFID and Deep Facial Biometry," in *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, Riyadh, Saudi Arabia, Apr. 2021, pp. 151–157.

[2] S. Dargan, M. Kumar, "A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities," *Expert Systems with Applications*, vol. 143, p. 113114, Apr. 2020.

[3] M. Turk, A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 71–86, Jan. 1991.

[4] P. Belhumeur, J. Hespanha, D. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711–720, July 1997.

[5] L. Wiskott, J.-M. Fellous, N. Kuiger, C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 775–779, July 1997.

[6] B. Hamdan, K. Mokhtar, "Face recognition using Angular Radial Transform," *Journal of King Saud University - Com- puter and Information Sciences*, vol. 30, pp. 141–151, Apr. 2018.

[7] R. Shoja Ghiass, O. Arandjelović, A. Bendada, X. Maldague, "Infrared face recognition: A comprehensive review of methodologies and databases," *Pattern Recognition*, vol. 47, pp. 2807–2824, Sept. 2014.

[8] Mamta, M. Hanmandlu, "Robust authentication using the unconstrained infrared face images," *Expert Systems with Applications*, vol. 41, pp. 6494–6511, Oct. 2014.

[9] M. Kanti Bhowmik, Kankan, S. Majumder, G. Majumder, A. Saha, A. Nath, D. Bhattacharjee, D. K. Basu, M. Nasipuri, "Thermal Infrared Face Recognition – A Biometric Identi- fication Technique for Robust Security system," in *Reviews, Refinements and New Ideas in Face Recognition*, P. Corcoran Ed., InTech, July 2011.

[10] M. Akhloufi, A. Bendada, J.-C. Batsale, "State of the art in infrared face recognition," *Quantitative InfraRed Thermography Journal*, vol. 5, pp. 3–26, June 2008.

[11] G. Pan, L. Sun, Z. Wu, S. Lao, "Eyeblink-based Anti-Spoofing in Face Recognition from a Generic Webcamera," in *2007 IEEE 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, 2007, pp. 1–8, IEEE.

[12] S. Jia, G. Guo, Z. Xu, "A survey on 3D mask presentation attack detection and countermeasures," *Pattern Recognition*, vol. 98, p. 107032, Feb. 2020.

[13] B. Hamdan, K. Mokhtar, "A self-immune to 3D masks attacks face recognition system," *Signal, Image and Video Processing*, vol. 12, pp. 1053–1060, Sept. 2018.

[14] S. Hu, N. Short, B. S. Riggan, M. Chasse, M. S. Sarfraz, "Het- erogeneous Face Recognition: Recent Advances in Infrared- to-Visible Matching," in *2017 12th IEEE International Con- ference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, DC, USA, May 2017, pp. 883–890, IEEE.

[15] R. Shoja Ghiass, H. Bendada, X. Maldague, "Université Laval Face Motion and Time-Lapse Video Database (UL-FMTV)," in *Proceedings of the 2018 International Conference on Quantitative InfraRed Thermography*, 2018, QIRT Council.

[16] T. Bourlai, A. Ross, C. Chen, L. Hornak, "A study on using mid-wave infrared images for face recognition," Baltimore, Maryland, May 2012, pp. 83711K–83711K–13.

[17] T. Bourlai Ed., *Face Recognition Across the Imaging Spectrum*. Cham: Springer International Publishing, 2016.

[18] S. G. Kong, J. Heo, F. Boughorbel, Y. Zheng, B. R. Abidi, Koschan, M. Yi, M. A. Abidi, "Multiscale Fusion of Visible and Thermal IR Images for Illumination-Invariant Face Recognition," *International Journal of Computer Vision*, vol. 71, pp. 215–233, Feb. 2007.

[19] D. Bhattacharjee, "Adaptive polar transform and fusion for human face image processing and evaluation," *Human- centric Computing and Information Sciences*, vol. 4, p. 4, Dec. 2014.

[20] A. R. Pal, A. Singha, "A comparative analysis of visual and thermal face image fusion based on different wavelet family," in *2017 International Conference on Innovations in Electronics, Signal Processing and Communication (IESC)*, Shillong, India, Apr. 2017, pp. 213–218, IEEE.

[21] G. Hermosilla, F. Gallardo, G. Farias, C. Martin, "Fusion of Visible and Thermal Descriptors Using Genetic Algo- rithms for Face Recognition Systems," *Sensors*, vol. 15, pp. 17944–17962, July 2015.

[22] N. K. Benamara, E. Zigh, T. Boudghene Stambouli, M. Keche, "Combined and Weighted Features for Robust Multispectral Face Recognition," in *Computational Intelligence and Its Applications*, vol. 522, 2018, pp. 549–560.

[23] N. K. Benamara, E. Zigh, T. B. Stambouli, M. Keche, "Efficient Multispectral Face Recognition using Random Feature Selection and PSO-SVM," in *Proceedings of the 2nd International Conference on Networking, Information Systems & Security - NISS19*, Rabat, Morocco, 2019, pp. 1–6.

[24] K. Guo, S. Wu, Y. Xu, "Face recognition using both visible light image and near-infrared image and a deep network," *CAAI Transactions on Intelligence Technology*, vol. 2, pp. 39–47, Mar. 2017.

[25] D. Lin, X. Tang, "Inter-modality Face Recognition," in *Computer Vision – ECCV 2006*, vol. 3954, A. Leonardis, H. Bischof, A. Pinz Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 13–26.

[4] https://github.com/nkbenamara/Terravic-Facial-IR-Database-Annotations-

[26] D. Yi, R. Liu, R. Chu, Z. Lei, S. Z. Li, "Face Matching Between Near Infrared and Visible Light Images," in *Advances in Biometrics*, vol. 4642, S.-W. Lee, S. Z. Li Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 523–530.

[27] F. Juefei-Xu, D. K. Pal, M. Savvides, "NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Boston, MA, USA, June 2015, pp. 141–150, IEEE.

[28] S. Liu, D. Yi, Z. Lei, S. Z. Li, "Heterogeneous face image matching using multi-scale features," in *2012 5th IAPR International Conference on Biometrics (ICB)*, New Delhi, India, Mar. 2012, pp. 79–84, IEEE.

[29] L. Huang, J. Lu, Y.-P. Tan, "Learning modality-invariant features for heterogeneous face recognition," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Nov. 2012, pp. 1683–1686.

[30] J. Lezama, Q. Qiu, G. Sapiro, "Not Afraid of the Dark: NIR-VIS Face Recognition via Cross-Spectral Hallucination and Low-Rank Embedding," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, July 2017, pp. 6807–6816, IEEE.

[31] K. Mallat, N. Damer, F. Boutros, A. Kuijper, J.-L. Dugelay, "Cross-spectrum thermal to visible face recognition based on cascaded image synthesis," in *2019 International Confer- ence on Biometrics (ICB)*, Crete, Greece, June 2019, pp. 1–8, IEEE.

[32] A. Kantarcı, H. K. Ekenel, "Thermal to Visible Face Recogni- tion Using Deep Autoencoders," *arXiv:2002.04219 [cs, eess]*, Feb. 2020.

[33] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Gener- ative Adversarial Networks," *arXiv:1406.2661 [cs, stat]*, June 2014.

[34] L. Song, M. Zhang, X. Wu, R. He, "Adversarial Discrimina- tive Heterogeneous Face Recognition," *arXiv:1709.03675 [cs]*, Sept. 2017.

[35] C. Chen, A. Ross, "Matching Thermal to Visible Face Images Using a Semantic-Guided Generative Adversarial Network," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, Lille, France, May 2019, pp. 1–8, IEEE.

[36] T. Zhang, A. Wiliem, S. Yang, B. Lovell, "TV-GAN: Gener- ative Adversarial Network Based Thermal to Visible Face Recognition," in *2018 International Conference on Biometrics (ICB)*, Gold Coast, QLD, Feb. 2018, pp. 174–181, IEEE.

[37] W.-T. Chu, P.-S. Huang, "Thermal Face Recognition Based on Multi-scale Image Synthesis," in *MultiMedia Modeling*, vol. 12572, J. Lokoč, T. Skopal, K. Schoeffmann, V. Mezaris, X. Li, S. Vrochidis, I. Patras Eds., Cham: Springer Interna- tional Publishing, 2021, pp. 99–110.

[38] X. Di, B. S. Riggan, S. Hu, N. J. Short, V. M. Patel, "Multi-Scale Thermal to Visible Face Verification via Attribute Guided Synthesis," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, pp. 266–280, Apr. 2021.

[39] H. Zhang, V. M. Patel, B. S. Riggan, S. Hu, "Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, Denver, CO, Oct. 2017, pp. 100–107, IEEE.

[40] B. S. Riggan, N. J. Short, S. Hu, "Thermal to Visible Synthesis of Face Images Using Multiple Regions," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, Mar. 2018, pp. 30–38, IEEE.

[41] Z.-Q. Zhao, P. Zheng, S.-T. Xu, X. Wu, "Object Detection With Deep Learning: A Review," *IEEE Transactions on Neu- ral Networks and Learning Systems*, vol. 30, pp. 3212–3232, Nov. 2019.

[42] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 580–587, IEEE.

[43] R. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1440–1448, IEEE.

[44] S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: To- wards Real-Time Object Detection with Region Proposal Networks," *arXiv:1506.01497 [cs]*, Jan. 2016.

[45] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016, pp. 779–788, IEEE.

[46] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, "SSD: Single Shot MultiBox Detector," in *Computer Vision – ECCV 2016*, vol. 9905, B. Leibe, J. Matas, N. Sebe, M. Welling Eds., Cham: Springer International Publishing, 2016, pp. 21–37.

[47] J. Redmon, A. Farhadi, "YOLOv3: An Incremental Improve- ment," *arXiv:1804.02767 [cs]*, Apr. 2018.

[48] A. Kumar, A. Kaur, M. Kumar, "Face detection techniques: a review," *Artificial Intelligence Review*, vol. 52, pp. 927–948, Aug. 2019.

[49] H. Jiang, E. Learned-Miller, "Face Detection with the Faster R-CNN," *arXiv:1606.03473 [cs]*, June 2016.

[50] R. Belaroussi, M. Milgram, "A comparative study on face detection and tracking algorithms," *Expert Systems with Applications*, vol. 39, pp. 7158–7164, June 2012.

[51] Y. K. Cheong, V. V. Yap, H. Nisar, "A novel face detection algorithm using thermal imaging," in *2014 IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE)*, Penang, Malaysia, Apr. 2014, pp. 208–213, IEEE.

[52] C. Ma, N. Trung, H. Uchiyama, H. Nagahara, A. Shimada, R.- Taniguchi, "Adapting Local Features for Face Detection in Thermal Image," *Sensors*, vol. 17, p. 2741, Nov. 2017.

[53] S. Yang, P. Luo, C. C. Loy, X. Tang, "WIDER FACE: A Face Detection Benchmark," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016, pp. 5525–5533, IEEE.

[54] M. Mirza, S. Osindero, "Conditional Generative Adversarial Nets," *arXiv:1411.1784 [cs, stat]*, Nov. 2014.

[55] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, "Image-to- Image Translation with Conditional Adversarial Networks," *arXiv:1611.07004 [cs]*, Nov. 2018.

[56] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, "Unpaired Image-to- Image Translation Using Cycle-Consistent Adversarial Net- works," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Oct. 2017, pp. 2242–2251, IEEE.

[57] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, S. P. Smolley, "Least Squares Generative Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Oct. 2017, pp. 2813–2821, IEEE.

[58] K. Simonyan, A. Zisserman, "Very Deep Convolutional Net- works for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Apr. 2015.

[59] O. M. Parkhi, A. Vedaldi, A. Zisserman, "Deep Face Recognition," in *Procedings of the British Machine Vision Conference 2015*, Swansea, 2015, pp. 41.1–41.12, British Machine Vision Association.

[60] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, Dec. 2015.

[61] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, "VGGFace2: A Dataset for Recognising Faces across Pose and Age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, Xi'an, May 2018, pp. 67–74, IEEE.

[62] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, N. Navab, J. Hornegger, W. M. Wells, F. Frangi Eds., Cham: Springer International Publishing, 2015, pp. 234–241.

[63] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Ra- jendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani, X. Yuan, "A Comprehensive Database for Benchmarking Imaging Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 509–520, Mar. 2020.

[64] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, X. Wang, "A Natural Visible and Infrared Facial Expression Database for Expression Recognition and Emotion Infer- ence," *IEEE Transactions on Multimedia*, vol. 12, pp. 682–691, Nov. 2010.

[65] S. Wang, Z. Liu, Z. Wang, G. Wu, P. Shen, S. He, X. Wang, "Analyses of a Multimodal Spontaneous Facial Expression Database," *IEEE Transactions on Affective Computing*, vol. 4, pp. 34–46, Jan. 2013.

[66] R. Pulecio, C. Gerardo, "Face recognition on distorted infrared images augmented by perceptual quiality-aware features," *S. Z. Li and A. K. Jain, Handbook of Face Recognition. New York, NY: Springer, 2005.*, Dec. 2016.

[67] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *International Journal of Computer Vision*, vol. 128, pp. 336–359, Feb. 2020.

[68] M. Kopaczka, K. Acar, D. Merhof, "Robust Facial Landmark Detection and Face Tracking in Thermal Infrared Images using Active Appearance

Models:," in *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Rome, Italy, 2016, pp. 150–158, SCITEPRESS - Science and Technology Publications.

[69] D. Poster, S. Hu, N. Nasrabadi, B. Riggan, "An Examination of Deep-Learning Based Landmark Detection Methods on Thermal Face Imagery," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, June 2019, pp. 980–987, IEEE.

[70] W.-T. Chu, Y.-H. Liu, "Thermal Facial Landmark Detection by Deep Multi-Task Learning," in *2019 IEEE 21st Interna- tional Workshop on Multimedia Signal Processing (MMSP)*, Kuala Lumpur, Malaysia, Sept. 2019, pp. 1–6, IEEE.

Nadir Kamel Benamara

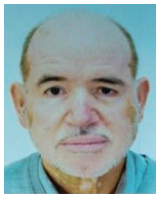
He received the "Ingenieur d'Etat" and the "Master" degrees in Advanced Telecom- munications Engineering from the Institut National des Télécommunications et des TIC d'Oran (INTTIC), Algeria in 2016. Currently, he is a PhD Candidate at the Université des Sciences et de la Technologie d'Oran - Mohamed Boudiaf (USTO-MB), Algeria. His research focuses on Biometrics, Computer Vision and Deep Learning.

Ehlem Zigh

She is an associate Professor class 'A' at Institut National des Télécommunications et des TIC d'Oran, Algeria. She received her Habilitation degree in Electronic field from Djillali Liabes University, Sidi Bel Abbès, Algeria in 2017. Her Doctorate degree from the University of Sciences and Technologies Mohammed Boudiaf of Oran, Algeria in 2014. She is a Head of a research group at LaRATIC Laboratory at Institut National des Télécommunica- tions et des TIC d'Oran (INTTIC), Algeria. She is a focal point of the Artificial Intelligence Training proposed by AI commons in Algeria. Her research interests include image processing, deep learning, soft computing techniques and internet of things. She has around 20 national and international communications, eleven international publications and she has published a book chapter in a Handbook of Research on Artificial Intelligence Techniques and Algorithms (DOI: 10.4018/978-1-4666-7258-1.ch010), Malaysia. Dr. E. Zigh has been a chairman session at the international SCA 2020 Online Conference. She is member of a technical program committee of EAI innovation research conferences. She is a reviewer at The International Journal of Energy Optimization and Engineering, IGI Global, Algerian journal of research and technology, European journal of remote sensing and Interactive Learning Environment journals.

Tarik Boudghene Stambouli

Doctorat d'Etat in Electronics (Option Signal Processing) from Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf – Department of Electronics, was graduated in Electronics. Lecturer in Department of Electrotech- nics, Member of Laboratoire Signaux et Images in same University, he currently focuses his research on biometrics.



Mokhtar Keche

Mokhtar Keche received the Ingenieur degree in Telecommunications from ENST Paris in 1978, and the Docteur Ingenieur degree and PhD from the University of Rennes in France and the University of Nottingham in U.K in 1982 and 1998, respectively. He is actually a Professor at the University of USTO in Algeria. His research interests are in the areas of Digital Communications, Array Processing, Multitarget Tracking, Road Traffic Estimation, and Biometry.