AI

MDPI

Review

Recent Advances in Infrared Face Analysis and Recognition With Deep Learning

Dorra Mahouachi [1] **and Moulay A. Akhloufi** [1,*]

Perception, Robotics, and Intelligent Machines Research Group (PRIME), Department Computer Science, University Moncton, Moncton, NB E1A 3E9, Canada
* Correspondence: moulay.akhloufi@umoncton.ca

Abstract: Besides the many advances made in the facial detection and recognition fields, face recognition applied to visual images (VIS-FR) has received increasing interest in recent years, especially in the field of communication, identity authentication, public safety and to address the risk of terrorism and crime. These systems however encounter important problems in the presence of variations in pose, expression, age, occlusion, disguise, and lighting as these factors significantly reduce the recognition accuracy. To prevent problems in the visible spectrum, several researchers have recommended the use of infrared images. This paper provides an updated overview of deep infrared (IR) approaches in face recognition (FR) and analysis. First, we present the most widely used databases, both public and private, and the various metrics and loss functions that have been proposed and used in deep infrared techniques. We then review deep face analysis and recognition/identification methods proposed in recent years. In this review, we show that infrared techniques have given interesting results for face recognition, solving some of the problems encountered with visible spectrum techniques. We finally identify some weaknesses of current infrared FR approaches as well as many future research directions to address the IR FR limitations.

Keywords: face recognition; deep learning; face analysis; feature extraction; infrared imaging; face synthesis

check for updates

Citation: Mahouachi, D.; Akhloufi, M.A. Recent Advances in Infrared Face Analysis and Recognition With Deep Learning. *AI* **2023**, *4*, 199–233. https://doi.org/10.3390/ai4010009

Academic Editors: Kenji Suzuki and José Manuel Ferreira Machado

Received: 21 December 2022
Revised: 16 January 2023
Accepted: 30 January 2023
Published: 7 February 2023

1. Introduction

In the last few years, deep learning researchers have shown increasing interest in the usage of face recognition (FR) since it is the main biometric technique for identity authentication. The development of FR tools has had an impact on the expansion of many areas such as the defense, finance, and service industries. Facial recognition uses computer-generated filters to transform face images into numerical values that can be compared to determine their similarity. This process consists of four steps:

- **Face detection**: Scanning the full image to identify whether or not the candidate area is a face.
- **Face preprocessing**: Performed on the detected area, which may consist of noise reduction, contrast enhancement, or similar operations.
- **Feature extraction**: The extraction of facial features such as eyes, nose, mouth, brows, and cheeks and the geometrical relation between them from the preprocessed facial image. In addition to Face recognition, the feature extraction step is used for emotion and pain detection.
- **Feature matching**: Use the extracted Feature vector to perform a comparison with a set of known faces.

Recent research has focused on the visible spectrum [1–6]. FaceNet [6] showed human-like performance after reaching an accuracy of 97.35% by training a 9-layer model on 4 million facial images therefore becoming the reference point for FR.

Despite the progress and results obtained in this field, FR remains a complex task as many factors may affect the results of the identity feature extraction including light, wearable accessories like hats and eyeglasses, facial expression, and head orientation [7–9]. Solving the poor illumination problem in images has become a necessity. Thanks to the availability of advanced infrared (IR) technology cameras, researchers have considered the use of IR images since they are less influenced by lighting in most of the normal operating conditions allowing for better results.

Many surveys have been conducted on the literature related to FR but most of them focus on visible FR (VIS-FR). These two surveys [9,10] presented a comprehensive review of classical infrared FR methods. Ouyang et al., Jin et al. and Dey et al. [11–13] discussed fusion techniques as researchers have found that combining visible and infrared techniques provides good performance rates and overcomes the limitations of both spectrums [9]. Kakkirala et al. [14] presented recent advances in thermal infrared face recognition and suggested new methods that could be explored to advance research.

1.1. Classic Face Recognition Methods

Research on the task of visible FR has made a major impact in the biometrics field by enabling the development of several holistic techniques. These techniques are currently used for infrared face recognition and include Principal Component Analysis (PCA) known as Eigenfaces [15], Linear Discriminant Analysis (LDA) [16], Independent Component Analysis (ICA) [17], Support Vector Machine (SVM) [18] and many other techniques.

The fusion of visible and infrared images has recently been used to achieve higher recognition results by merging either images or results. This process is used to maximise the amount of useful information that is gathered from various images and to convert all images into a single image. In this context, Heo et al. [19] have shown the efficiency of these techniques by achieving high results. This fusion approach solved several issues as shown in the work of Chen et al. [20]. The authors used a decision-based scheme and a fuzzy integral [21] to merge the objective evidence provided by each modality which considerably improved recognition performance. In the work of Akhloufi et al. [22], a fusion framework operating in the infrared spectrum has been developed. Both active and passive infrared modalities are used in this framework. The proposed method uses intra- and inter-spectral fusion and operates in texture space. Compared to non-fused images, multi-scale fusion techniques (pyramidal and wavelet-based) improve the recognition performance.

The sensitivity of IR thermal images to the facial occlusion caused by eyeglasses was discussed by Bebis et al. [23]. The authors suggested combining IR and visible imaging to get around this issue. Since the features of the faces are captured differently by IR and visible imaging, a more accurate description of the face might be found by combining the complementary data from the two spectra. Also, Kong et al. [24] used an elliptical fitting method to find the location and shape of eyeglasses. They then replaced the eyeglass regions with an average thermal eye template to reduce the eyeglasses' effect. This technique helped confirm the study of Bebis et al. [23]. Most of the Fusion-based methods used FaceIt [25] (a commercial face recognition software) as a single recognition module.

For multispectral face identification, non-linear learning and subspace recognition techniques have been proposed by Akhloufi et al. [26]. The performance of global and local non-linear approaches was compared to the performance of traditional linear techniques. Undesirable variations caused by illumination changes, facial emotions, posture, and other factors can be avoided or reduced by using non-linear methods.

Holistic techniques aim to discriminate features that are only related to the identity of the face and to ignore domain information. LBP (local binary pattern) [27] is one of the most common infrared face recognition techniques. The facial area is first divided into small regions from which local binary patterns (LBP) are extracted. The histograms are then concatenated into a single vector. This vector provides an efficient representation of the face which is used to calculate the similarity between images. In the work of Li et al. [28], the authors presented two statistical learning methods for face recognition invariant to indoor lighting using NIR images. With the goal of building face recognition classifiers from a variety of LBP features, they used LDA [29] and AdaBoost [30] to achieve a high accuracy face recognition engine. Furthermore, in the work of Mendez et al. [31], the merits of the Local Binary Pattern (LBP) representation are studied in the context of face recognition using long-wave infrared images. These images are invariant to lighting, but at the same time they are affected by the fixed background noise inherent to this technology. The fixed pattern is normally compensated for using a non-uniformity correction method. This study shows that the LBP approach performs well under fixed noise and in the presence of glasses. No noise suppression pre-processing was required, however, if a non-uniformity correction method is applied, the image texture is amplified and the performance of LBP is degraded. The application of this approach as texture descriptors for efficient multispectral face recognition was presented by Akhloufi et al. [32]. The success rate of texture identification algorithms exceeded that of untransformed images, especially in the infrared spectrum. The effects of noise, light change, and facial expression are less significant in the suggested texture space. For this single technique (LBP), we can cite several related methods. Huang et al. [33] used ELBP (Extended Local Binary Pattern) [34] for face recognition in near-infrared (NIR) lighting conditions to solve the issues caused by lighting variations. Also, Zhao et al. [35] presented an illumination invariant dynamic facial expression recognition in NIR video sequences. The LBP-TOP feature descriptor (local binary patterns of three orthogonal planes) [36] can describe appearance and movement and is invariant to monotone grayscale changes. Zhao et al. used it on NIR images to provide an accurate result for video-based facial expression recognition with an invariant illumination system. Another method used in the work of Xie et al. [37] is the LBP co-occurrence matrix [38]. Xie et al. presented a new method of IR face recognition based on the LBP co-occurrence matrix to extract spatial relations between LBPs to describe the infrared face as traditional LBP-based feature fails to consider space location information.

SIFT (Scale-invariant Transform function) [39] is a face recognition algorithm that uses computer vision techniques to detect and define the spatial features of images. The key points of SIFT objects are first extracted from a collection of reference images and stored in a dataset. Faces are then recognized in a new image by individually comparing each attribute of the current image to the dataset and by identifying similar features for the candidates according to the Euclidean distance of their vectors. Yang et al. [40] presented a partial face alignment method based on the Scale Invariant Characteristic Transform (SIFT). First, a reference model is trained using holistic faces where anchor points and their corresponding descriptor subspaces are learned from the initial key points of the SIFT. The relationships between the anchor points are also extracted. Then, for the alignment, they used a mapping between the partial face key points and the anchor points of the learned face model to match the learned holistic face model to an input partial face image. To eliminate outlier correspondence, a shape constraint is used in this case and a temporal constraint is applied to find more outliers. Alignment is finally achieved by solving a similarity transformation.

In addition to the various techniques presented above, Zou et al. [41] proposed a new method presented under the name of active near-infrared lighting to overcome the illumination problem. This method uses an LED light source to provide a constant invisible lighting condition by enhancing the automatic detection of the eye through this light pupil. For experiments on such data, they used the Weka Machine Learning Toolkit [42]. In the work of Friedrich et al. [43], they focused on researching the effect of pose and facial expressions in FR using IR images showing that their results are less affected by these factors than with visible images. In another context, Wu et al. [44] tried to find a way to improve the performance of IR face recognition in different environments. They used, for this paper, blood perfusion rates that were obtained from the distribution of appearance temperature and they considered that these physiological characteristics are invariant to changes. In the infrared spectrum, Akhloufi et al. [45] proposed a method for extracting facial physiological features. The network of blood vessels under the skin is represented by these features. They used a distance transform to obtain an invariant representation for face recognition [45]. The extracted physiological features are related to the location of blood vessels under the skin of the face. Each person's blood network is unique, and it can be used in infrared face recognition.

1.2. Contributions and Outline

Our paper aims to provide a comprehensive review of recent advances in IR facial analysis and recognition research, particularly for deep learning methods. Most of the suggested work is focused on deep convolution neural networks (CNN) and synthesis techniques, especially heterogeneous techniques. CNN are a class of neural networks programmed to learn the parameters of the convolution from a collection of available data during training. They are composed of multiple layers, such as convolution layers, deconvolution layers, pooling layers, and so on. In recent years, new architectures have been proposed to improve performance and solve some traditional CNN limitations. CNN are one of the most common deep learning methods in face analysis. Image synthesis for machine learning applications provides the means to produce vast volumes of training data effectively. Synthetic data can become a critical component of the training pipeline of deep learning applications. Many training approaches for producing data have been developed during the last decade.

In the following, we will present the main datasets used for IR Face recognition research in Section 2 and we will introduce the different metrics and loss functions in Sections 3 and 4. Common techniques used to assess deep learning models are presented in Section 5. The techniques are divided into three categories: Synthesis methods (Section 5.1), Feature learning methods (Section 5.2), and NIR-VIS alignment methods (Section 5.3). Some applications are also introduced (Section 5.4). Comparative analysis of the different algorithms is presented and discussed in Section 6.

2. IR Datasets

Data is a critical asset for many research fields. This is particularly true for facial images for face recognition research. For infrared face recognition, the first problem encountered is the availability of accessible infrared datasets. This is not the case for visible spectrum datasets as multiple are available [46–52]. Compared to the datasets available for visible face recognition, we can notice the lack of resources for IR face recognition. Table 1 gives a summary of the IR face recognition datasets.

Table 1. IR FR Datasets (HO: Head Orientation, FE: Facial Expression).

Dataset	#Images	#Subjects	Accessories	Variations	Spectrum
CASIA [53]	3940	197	glasses	HO, FE	NIR
PolyU [54]	3500	350	-	HO, FE	VIS, NIR
USTC-NVIE [55]	-	215	glasses	HO, FE	VIS, thermal
Oulu-CASIA [35]		80	-	FE	VIS, NIR
IRIStcite [56]	4190	30	-	HO, FE	Thermal
CSIST [57]	1000	50	-	-	VIS, NIR
UL-FMTV [58]	-	238	glasses	HO, FE	Thermal
High-Resolution Thermal Face Dataset [59]	300	30	glasses	HO, FE	Thermal
Fully Annotated Thermal Face dataset [60]	2500	90	-	HO, FE	Thermal
RGB-D-T [61]	45,900	51	-	HO, FE	VIS, thermal
HIT LAB2 [57]	2000	50	-	HO, FE	VIS, NIR
SunWin [62]	4000	100	-	HO, FE	VIS, NIR
University of Notre Dame's UND collection X1 [63]	4584	82	-	HO, FE	VIS, LWIR
μ-faces dataset [64]	11,660	35	glasses	HO, FE	VIS, NIR, MWIR, LWIR
ARLV-TF [65]	500,000	395	glasses	HO, FE	VIS, LWIR
BUAA-VIS-NIR [66]	2700	150	-	HO, FE	VIS, NIR
ND-NIVL [67]	24,605	574	-	-	VIS, NIR
Polarimetric thermal dataset [68]	800	60	-	HO, FE	VIS, LWIR
SC3000-DB [69]	766	40	-	-	NIR
CARL [70]	7380	41	-	-	VIS, Thermal, NIR
Terravic [71]	-	20	glasses	HO, FE	Thermal
The IIIT Delhi occluded dataset [72]	1362	75	multiple	HO, FE	VIS, Thermal
INF [73]	470	94	-	-	NIR
TUFTS [74]	10,000	113	glasses	HO, FE	VIS, NIR, Thermal
Charlotte-ThermalFace database [75]	1000	10	-	HO, FE	Thermal

2.1. CASIA NIR Dataset

The CASIA NIR dataset [53] contains 3940 images of 197 subjects with a resolution of 640 × 480 pixels. Figure 1 shows the variations in head orientation, facial expressions, and wearable accessories that are present in this dataset. The images were captured in an environment using a NIR light-emitting diode (LED) as an active radiation source. To allow the NIR light to pass through, they blocked the visible light using a long-pass optical filter. For this dataset, they used a custom camera with a wavelength of 850 μm.

2.2. PolyU NIR Face Dataset

The PolyU-NIRFD [54] dataset includes images of 350 subjects, each appearing in 100 images at a resolution of 768 × 576 pixels. It contains NIR and visible images showing different head positions and facial expressions with scale variations and time intervals. A JAI camera with a wavelength of 850 μm was used to collect the images.

2.3. USTC-NVIE Dataset

The USTC-NVIE dataset [55] includes 2 sets of 215 subjects, spontaneous and posed, containing infrared and visible images with different orientations (frontal, left, right), facial expressions, and glasses (See Figure 2). To obtain the images, an SAT-HY6850 infrared camera capturing 25 frames per second with a resolution of 320 × 240 pixels and wave bands of 8–14 μm was used.

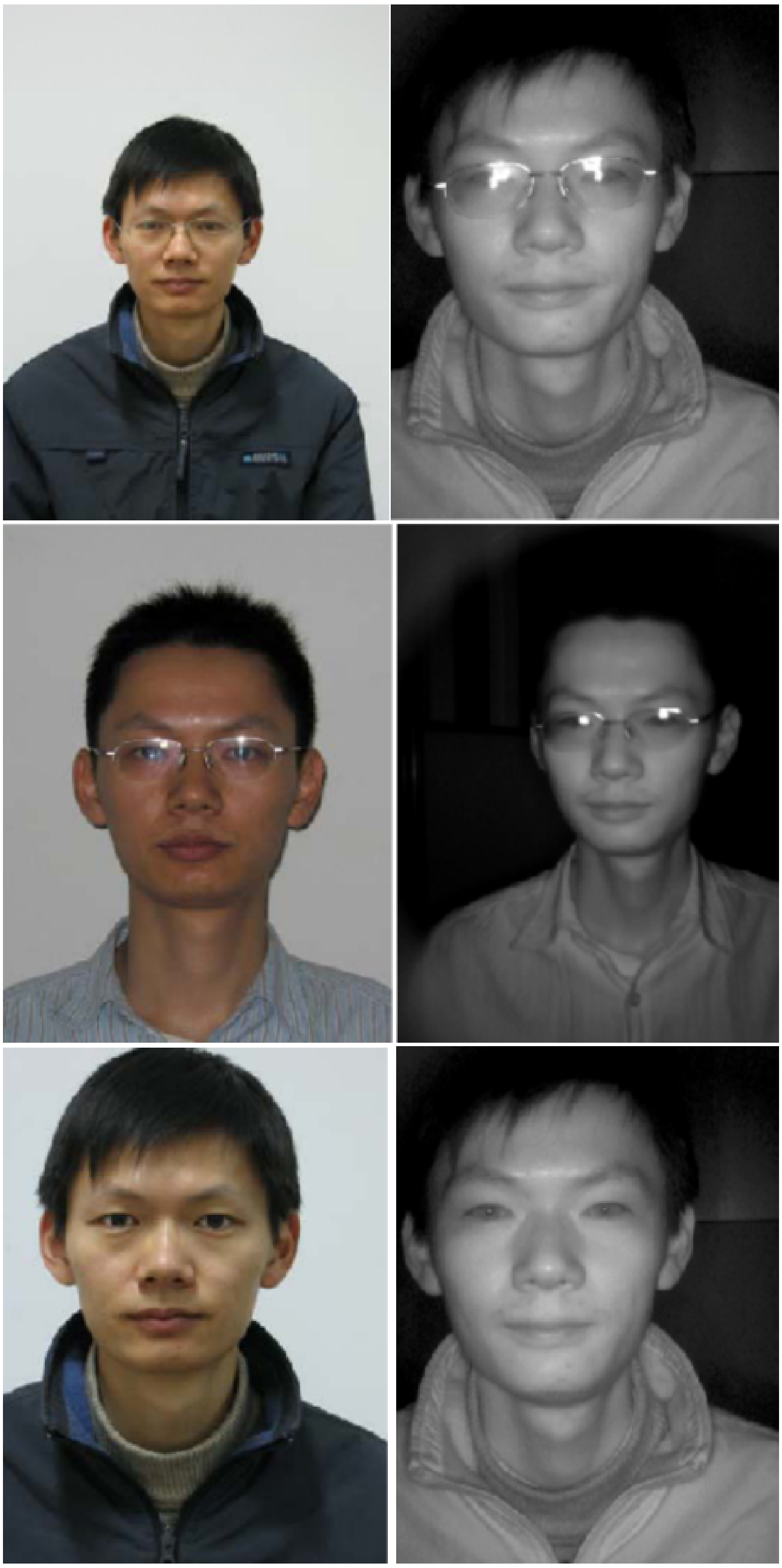

Figure 1. VIS and NIR face images, with variations in resolution, lighting conditions, pose, and age, of one subject in the NIR-VIS2.0 dataset [53].

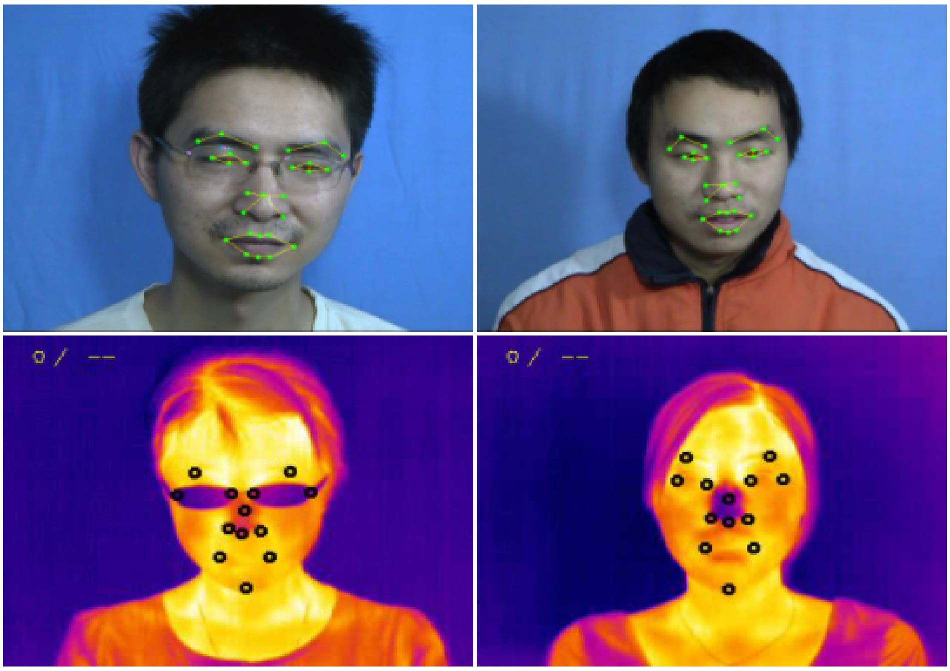

Figure 2. VIS and NIR face images with and without glasses [55].

2.4. Oulu-CASIA NIR-VIS Dataset

The Oulu-CASIA [35] dataset consists of 80 subjects aged 23–58 years (73.8% of subjects are male). Six variations of expression were considered while capturing the dataset: anger, disgust, fear, happiness, sadness, and surprise. The dataset consists of two parts, one taken by the Oulu University artificial vision group, comprising 50 subjects, mostly Finns. The other part was taken in Beijing by the Chinese Academy of Sciences National Pattern Recognition Laboratory and includes 30 subjects, all Chinese. The subjects were facing the camera. A USB 2.0 cameras for PC (SN9C201 & 202) including an NIR sensor and a VIS camera was used to capture the same facial expression with an image resolution of 320 × 240 pixels.

2.5. IRIS Dataset

The IRIS dataset [56] consists of images collected from 30 people totaling 4190 thermal images. Subjects performed subtle head movements. On average, 11 heads pose images per subject are presented. The dataset captures three facial expressions: sad, surprised, and laughing. The IRIS dataset was recorded with the Raytheon Palm-IR-Pro-camera with a spatial resolution of 320 × 240 pixels.

2.6. CSIST Dataset

The Harbin Institute of Technology Shenzhen Graduate School published the CSIST dataset [57]. It contains facial images captured in various lighting environments. It has two main datasets: Lab1 and Lab2. The Lab1 dataset includes 500 visible and 500 NIR images of 50 subjects at a resolution of 100 × 80 pixels and the Lab2 includes 1000 visible and 1000 NIR images of 50 subjects at a resolution of 200 × 200 pixels.

2.7. UL-FMTV

The dataset [58] consists of 238 subjects divided in two categories, genuine and impostor, composed of 134 subjects and 104 subjects, respectively. The face photos were taken between 2010 and 2014 to have a period of two to four years between the face photo sessions for most subjects. Pose, facial expressions, ethnicity, aging, time lapse, and eyeglass opacity are all considered in this dataset (See Figure 3). The researchers used a high-resolution

Indigo Phoenix thermal camera to capture the dataset and provided an image resolution of 640 × 512 pixels.

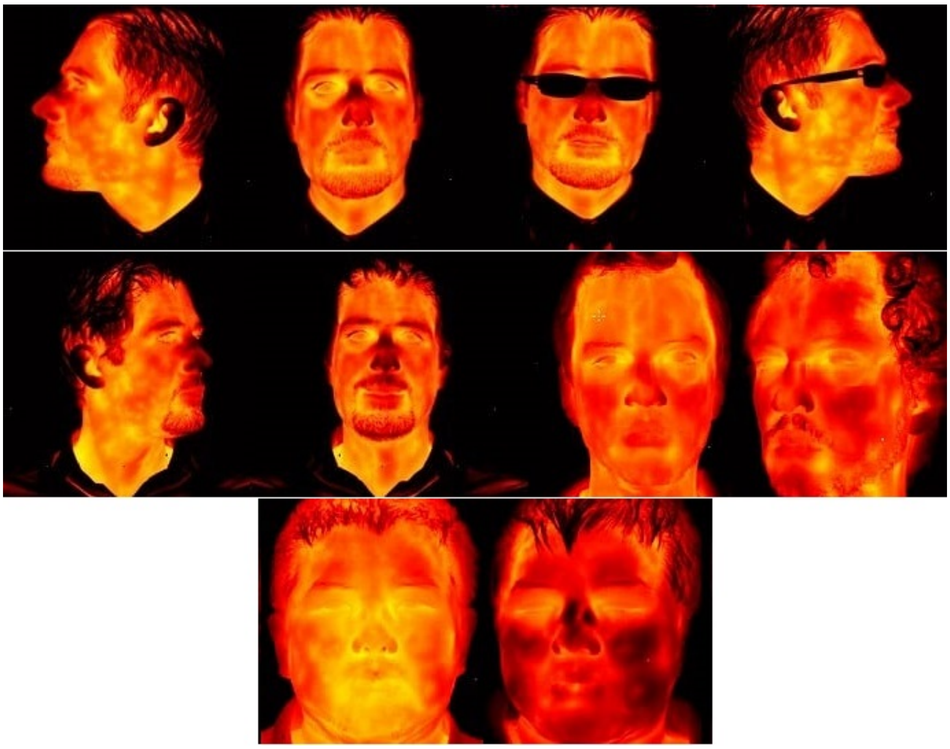

Figure 3. A sample of UL-FMTV after a time lapse with different head orientations, glasses and temperature exposure [58].

2.8. RGB-D-T Face Dataset

This dataset [61] is composed of 45,900 images from 51 subjects, mostly white males between the ages of 20 and 40. Their faces were captured under different conditions of movement, facial expressions, and lighting. A Microsoft Kinect for Windows was used to capture RGB and depth images and an AXIS Q1922 sensor was used to capture thermal images. The resolutions of the RGB, depth and thermal images are 640 × 480, 640 × 480, and 384 × 288 pixels, respectively.

2.9. ND-NIVL

The dataset [67] contains images of 574 subjects obtained between fall 2011 and spring 2012 that are visible light and near-infrared images. 2341 visible light facial images and 22,264 near-infrared facial images are present in this dataset. At least 402 participants had both visible and near-infrared images. The NIR images and the visible images have a resolution of 4770 × 3177 pixels and 4288 × 2848 pixels respectively. This makes ND-NIVL the largest database of high-resolution NIR and VIS images.

2.10. CARL Dataset

For the CARL dataset [70], visible and thermal images were acquired using a TESTO 880-3 thermographic camera equipped with an un-cooled detector with spectral sensitivity ranging between 8 and 14 µm. For the near-infrared, a customized Logitech Quickcam E2500 messenger was used, equipped with a silicon-based CMOS image sensor with sensitivity for the entire visible spectrum and half of the near-infrared. The thermographic camera offers a resolution of 160 × 120 pixels for thermal images and 640 × 480 for visible images and the webcam offers a maximum resolution of 640 × 480 pixels for near-infrared images. The dataset includes 41 subjects: 32 men and 9 women. Each person participated in four acquisition phases and provided five shots in three lighting conditions (See Figure 4) totaling 7380 images.

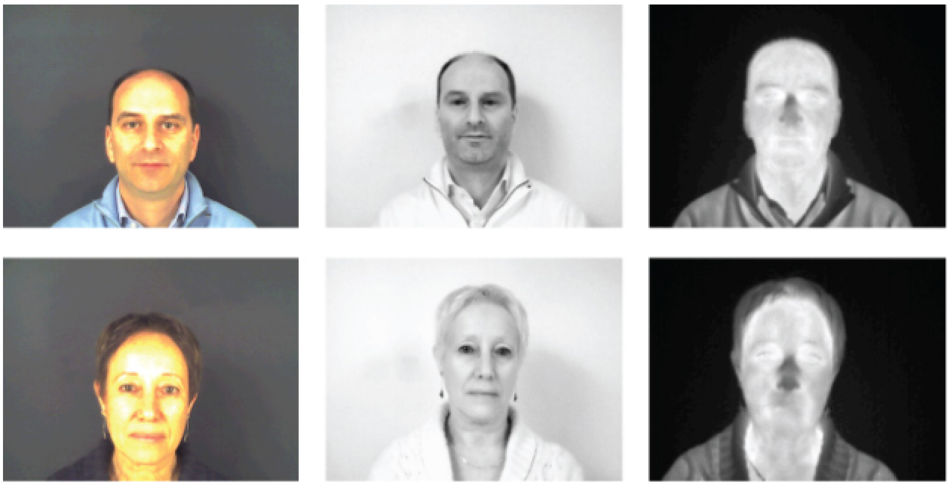

Figure 4. Example images from Carl dataset [70].

2.11. *University of Notre Dame's UND Collection X1*

The dataset [63] was assembled using an un-cooled LWIR sensor from Merlin and a high-resolution visible color camera. The resolution of the visible images is 1600 × 1200 pixels and the resolution of the thermal images obtained by LWIR is 312 × 239 pixels. The collection includes images in three experimental settings: expression (neutral, smile, laugh), light change, and time lapse. The data set includes 4584 images of 82 subjects in the visible and thermal range.

2.12. *μFaces Dataset*

μ-faces is a multispectral database used to conduct experimental tests of face recognition presented in [64]. This dataset contains the following spectra: visible, near-infrared, mid-infrared, and long-wave infrared. The dataset consists of 35 individuals and 11,660 images. The camera allows direct acquisition of visible and NIR images at a resolution of 640 × 480 pixels. The resolution of the MWIR camera is 640 × 512 pixels, and the resolution in LWIR is 160 × 128 pixels. This dataset contains faces in various scenarios with varying facial expressions, glasses, varying time and metabolic changes.

2.13. *ARLV-TF Dataset*

The DEVCOM Army Research Laboratory visible and thermal face (ARLVTF) was collected by Poster et al. [65]. The ARL-VTF dataset is the largest collection of matched visible and thermal facial images to date, with over 500,000 images of 395 individuals. A modern long-wave infrared (LWIR) camera was installed alongside a stereo arrangement of three visible spectrum cameras to collect the data. The camera gives an image resolution of 640 × 512. Variations of expression, head poses, and eyeglasses have been carefully captured in order to replicate real-world situations.

2.14. *UNC Charlotte Thermal Face Database*

The UNC Charlotte Thermal Face is the first publicly available new thermal database [75]. The Charlotte-ThermalFace database contains more than 10,000 infrared thermal images of 10 healthy subjects under different thermal conditions, at different distances from the camera, and with different head positions. It contains an annotation of the thermal sensation of each subject under different thermal conditions. The images were taken at 10 different distances from the camera for each temperature range. They used the FlirA7000 for this purpose. The original resolution of the thermal sensor is 640 × 480 pixels, and the resolution of the cropped facial area varies for each distance range.

2.15. Small Datasets

On top of the public datasets presented in the previous sections, less known and smaller datasets are available. We can present the HIT LAB2 face dataset [57] that contains 2000 face images of 50 volunteers. The dimensions of the images are 200 × 200 pixels. These images were taken under special lighting conditions. The images often show significant changes in posture or facial expression. The dataset is collected and distributed by the Harbin Institute of Technology. The BUAA-VIS-NIR face dataset [66] contains images of 150 subjects, with 9 VIS and 9 NIR images taken simultaneously for each subject. The nine images of each subject correspond to the neutral expression frontal, left rotation, right rotation, upward tilt, downward tilt, joy, rage, sadness, and surprise. Hu et al. proposed the polarimetric Thermal dataset [68]. It contains polarimetric LWIR images and visible spectrum images of 60 distinct subject's collections. This dataset was acquired using a polarimetric long-wave infrared imager, specifically a division-of-time spinning achromatic retarder system, which acquires geometric and textural details of the face that are not available in conventional thermal imaging. Subjects were asked to count out loud from 1 to 10 to capture mouth movements and, to some extent, to produce variations in facial imaging. Five hundred images are recorded with the polarimeter, and 300 images are recorded with a visible spectrum camera. Ariffin et al. [71] proposed the terravic Face IR Dataset composed of 20 individuals, including for each them various combinations (front, left, right, indoor/outdoor, hat) as shown in Figure 5. The format of the images is an 8-bit grayscale JPEG with a resolution of 320 × 240 pixels. This dataset was captured using Raytheon L-3 Thermal-Eye 2000AS. The IIIT Delhi occluded dataset [72] contains visible and thermal spectrum images of 75 participants with disguise variations. The dataset has 6 to 10 images per individual. There is at least one frontal neutral face image and at least five frontals disguised face images for each individual. There are 681 images for each spectrum. A thermal camera with a micro-bolometer sensor operating at 8–14 µm was used to collect the thermal images. The thermal images have a resolution of 720 × 576 pixels. All the images of the faces were taken with consistent lighting, neutral expressions, and in a frontal position. The dataset's images presents disguise variations such as variations due to different hairstyles and to the presence of beards and mustaches, glasses, caps and hats, and masks.

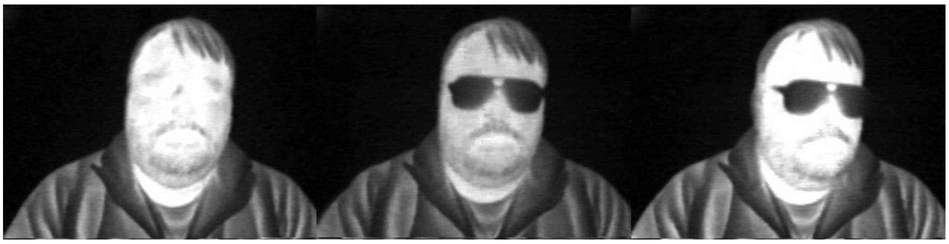

Figure 5. A sample of Terravic dataset [71].

2.16. Private Datasets

Besides the datasets presented above, many others have been created for the same purpose of infrared face recognition. The high-resolution thermal face dataset for Face and Expression Recognition [59] and the fully annotated thermal face dataset [60] are both private datasets. The first is divided into 2 sets of images containing 30 subsets. The subsets each contains 30 subjects, including 10 image sets of subjects with glasses collected over 12 months. The images show subjects with different head positions and expressions. The second dataset was developed from 2500 high-resolution (1024 × 768 pixels) fully annotated images of 90 subjects. It contains various presentations of head poses and facial expressions such as basic morphological changes, fundamental emotions, and arbitrary expressions. To obtain this dataset, the images were taken with a high-resolution thermal infrared camera with a 30 mm f/1.0 prime lens Infratec HD820.

Another private dataset to mention is the Sunwin dataset [62]. It features 4000 images of 100 subject faces. This data set is divided into two subsets; the first subset contains 2000 visible-light images from the 100 subjects. The second set contains 2000 near-infrared images of the 100 subjects. For both sets, 10 images are taken under normal light for each person, and the remaining 10 images are taken under abnormal light. The extracted collection includes various facial expressions, lighting, and other changes. Data was collected using a visible light camera and a near-infrared camera at the same time.

Szankin et al. [69] created the SC3000-DB dataset while studying the influence of thermal imagery resolution on the accuracy of deep learning. The dataset was created using a FLIR ThermaCam SC3000 infrared camera with a resolution of 320 × 240 pixels in a noise reduction mode. It contains 766 images of 40 volunteers from a cohort of 19 males and 21 females.

Other datasets used in the papers presented below that are important to cite include the INF dataset [73] and TUFTs dataset [74]. The INF dataset consists of 470 near-infrared images taken by a near-infrared camera of 94 subjects. Each participant has five NIR facial images with a resolution of 640 × 480 pixels. TUFTS is the most complete, large-scale face dataset available, with over 10,000 pictures taken from 74 females and 39 males from over 15 countries, ages 4 to 70, and with six image modalities: visible, near-infrared, thermal, computerized sketch, recorded video, and 3D images. A large-scale thermal facial database is included in the Tufts face dataset with pose variance and facial expressions.

3. Metrics

We must keep in mind that for any non-trivial problem, no machine learning algorithm is perfect. It is therefore crucial to evaluate the performance of the algorithm in order to adapt it to an application. Many metrics have been developed to evaluate face recognition methods in the visible and infrared spectrum. In this section, we mention the most well-known metrics used for infrared FR.

3.1. Receiver Operating Characteristic (ROC)

The ROC (Receiver Operating Characteristic) approach [76], which was originally used in the early 1980s, has since become a widely used method for assessing detection performance. It was first used to assess the diagnostic capabilities of medical imaging systems, especially in radiology. The ROC analysis for a single-target problem involves the true positive rate tpr and the false positive rate fpr. These parameters are calculated in the ROC analysis to identify the binary response of the detection system to a stimulus (an image), with:

$$tpr = \frac{TP}{total\ positives} \tag{1}$$

$$fpr = \frac{FP}{total\ positives} \tag{2}$$

A point in the ROC plane is represented by the pair (fpr;tpr). ROC curves are created by altering the detection system's parameters and calculating tpr and fpr for each value of the fixed parameters. ROC analysis is an excellent approach for assessing detection performance since it accounts for the rate of each class and offers two direct antagonistic measurements that are crucial to the system's calibration.

3.2. Mean Accuracy (ACC)

The accuracy (Acc) [77] is computed by dividing the total number of correct predictions by the total number of data points in the dataset. The highest value of accuracy is 1.0 and the lowest is 0.0. The accuracy is the probability of properly classifying a random example and it is correlated to the actual and predicted classes. It is mathematically defined by the equation below:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

where TP, TN, FP and FN represent true positives, true negatives, false positives and false negatives respectively.

3.3. Validation Rate (VAL) and False Accept Rate (FAR)

The validation rate is proportional to the number of times the same person's face was properly identified. The false acceptance rate is proportional to the number of times two separate people's faces have been mistakenly recognized as being the same individual [6]. P_{same} denotes all pairs of faces (i,j) having the same identity, whereas P_{diff} denotes all pairings of differing identities. For a given face distance d, the validation rate $VAL(d)$ and the false acceptance rate $FAR(d)$ are defined as:

$$VAL(d) = \frac{TP(d)}{Psame} \tag{4}$$

$$FAR(d) = \frac{FP(d)}{Pdiff} \tag{5}$$

where $TP(d)$ and $FP(d)$ represent true positives and false positives respectively.

3.4. Cumulative Matching Characteristics (CMC)

The cumulative matching characteristics [78] are used for assessing the measured accuracy performance of a biometric system as it performs identification in a closed environment. The templates are compared and ranked based on how similar they are. Based on the match rate, the CMC indicates how frequently the biometric subject template appears in the ranks (1, 5, 10, 100, etc.). The identification rate is compared to the rank (1, 5, 10, 100, etc.) in a CMC.

3.5. Precision-Coverage Curve

Precision-coverage curves are used to assess the accuracy of identification under a changing threshold t. When the probe's confidence score falls below t, it is rejected. The algorithms are compared in terms of the proportion of probes passed (or coverage) and a high recognition accuracy is around 95% or 99%.

3.6. Minimum Squared Error (MSE)

For pattern recognition, the Minimum Squared Error (MSE) method [79] is commonly employed and it performs well for face classification. In terms of classification, MSE has many advantages. It is straightforward and simple to implement. The MSE approach may be used not only for two-class classification but also for multi-class classification. The MSE algorithm utilizes the training sample and its class label from the training mapping to predict the test sample's class label. It then uses the resulting mapping to predict the test sample's class label. MSE then selects the training sample that is the most similar to the test sample. Finally, MSE classifies the test sample within the the training sample's class.

4. Loss Functions

The loss function used by a model is arguably the most important element in determining its performance. Algorithms learn by means of a loss function. They determine how effectively a particular algorithm models the data. The loss function will provide a very high value if the predictions diverge too far from the actual results. The loss function gradually learns to minimize the prediction error with the help of an optimization function. Several loss functions used in IR FR will be discussed in the following.

4.1. Softmax Loss

The categorical cross-entropy loss with Softmax activation in the final layer is known as Softmax loss. In multi-class classification problems, categorical cross-entropy is a common loss function. This is a task in which an example may only fit into one of several categories and the model must choose which one to classify them in. Its formal purpose is to calculate the difference between two probability distributions. The softmax activation resizes the model's output to give it the proper properties. The following is how the softmax loss is defined:

$$L_S = -\sum_{i=1}^{m} \ln \frac{\exp\left\{W_{y_i}^T x_i + b_{y_i}\right\}}{\sum_{j=1}^{n} \exp\left\{W_{y_j}^T x_i + b_j\right\}} \tag{6}$$

The feature vector of the i^{th} image is defined as x_i. The j^{th} column of weights is defined as W_j, and the bias term as b_j. The number of classes and images is n and m, respectively, with y_i denoting the class of the i^{th} image.

4.2. Triplet Loss

The triplet loss is a loss function that focuses on reducing the distance between an anchor (an image of a person) and positive example data (another image of the same person). Faces with the same identification should look closer to each other than faces with different identities. The triplet loss function is used to train the neural network's parameters so that it can encode images properly. Comparing pairs of pictures is the purpose of this function. As inputs, three images are used: an anchor image, a positive image, and an image of a different person is used as a negative image.

$$L = \frac{1}{N} \sum_{i}^{N} \max\left(\|f(A_i) - f(P_i)\|^2 - \|f(A_i) - f(N_i)\|^2 + m, 0 \right) \tag{7}$$

Anchors, positive examples, and negative example pictures are represented as A_i, P_i,and N_i, respectively. The embedding of these pictures in the feature space are represented as $f(A_i)$, $f(P_i)$, and $f(N_i)$. m is the profit margin. N is the cardinality of all possible triplets in the training set.

4.3. Center Loss

The authors in [80] came up with the notion of central loss to overcome the limitations of Softmax loss. First, they discovered that the distribution of data in the feature space has a lot of intra-class variance. They illustrate this with a model with only two fully connected nodes in the final layer. To address this issue, they included an extra term to the softmax loss that penalizes the model if the data points are distant from the class centroid:

$$L_c = \frac{1}{2} \sum_{i=1}^{m} \left\| x_i - c_{y_i} \right\|_2^2 \tag{8}$$

$$L = L_S + \lambda L_c \tag{9}$$

The softmax loss is denoted by the symbol L_S. The centroid of all points corresponding to the y_i class of the i^{th} data point in feature space is c_{y_i} in L_c. L_c is the total of all point's squared distances from their respective class' centroid. The size of the mini-batch is defined as m. Instead of computing the centroid for the full data set, it is computed for each batch separately. The circle loss explicitly penalizes intra-class variation.

4.4. Mutual Component Analysis Loss

For mutual component extraction, the original MCA uses pairs of pictures with the same identity for its input. However, because we want to conduct offline feature extraction in real-world applications, feature extraction from a single picture is more convenient for computer vision applications. Therefore, in [81] they propose the MCA loss which enforces each modal-dependent component, i.e., $\vec{E^{1,k}_{M_i}}$ to approach the mutual component, i.e., $\vec{E^{1}_{M_i}}$. In other words:

$$\vec{E^{1,1}_{M_i}} = \vec{E^{1,2}_{M_i}} = \vec{E^{1}_{M_i}} \tag{10}$$

which can be written as

$$\vec{E^{1}_{M_i}} = \frac{1}{2}\left(\vec{E^{1,1}_{M_i}} + \vec{E^{1,2}_{M_i}}\right) = \vec{E^{1,1}_{M_i}} = \vec{E^{1,2}_{M_i}} \tag{11}$$

The modal discrepancy is fully removed in this perfect situation, as shown by this equation. However, there is no proof that this is always the case. As a result, they attempt to keep the gap between $\vec{E^{1,k}_{M_i}}$ and $\vec{E^{1}_{M_i}}$ as small as possible. As a result, the MCA loss may be expressed as

$$L_{mca} = \frac{1}{2NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\left\|\vec{E^{1,k}_{M_i}} - \vec{E^{1}_{M_i}}\right\|_2^2 \tag{12}$$

where K denotes the number of modalities and N is the number of training images in modality k. They feed image pairs into the network so the total number of images is $N \times K$.

4.5. Modality Discrepancy Loss

The MD loss, introduced in [82], reduces modal disagreement by reducing the cosine distance between modalities. Given that we generally utilize cosine similarity to determine the difference between two face pictures, Deng et al. used the cosine distance as $diff(*,*)$. To that aim, the modality discrepancy loss (MD loss) between ($\hat{X}^v_i$,$\hat{X}^n_i$) is defined as follows:

$$L_{MD} = \frac{1}{N}\sum_{i=1}^{N}(1 - \cos(\hat{X}^v_i, \hat{X}^n_i)) \tag{13}$$

The cosine similarity of two inputs is $cos(*,*)$ and the total number of picture pairings is N. By optimizing L_{MD}, it forces two facial representations to be similar. The entire loss may be expressed as follows:

$$L = L_S + \lambda L_{MD} \tag{14}$$

where L_S is the facial classification cross entropy loss and λ is a hyper-parameter that acts as a trade off these two components.

4.6. Component Adaptive Triplet Loss

Xu et al. [83] proposed a component adaptive triplet loss function (L_{CAT}) that takes into account changes in pose or emotion by assigning adaptive weights based on the visible region of the face. The positive example is a picture with the same ID as the anchor example and the negative example is a picture with a different ID. The authors sampled anchor and positive into distinct domains and negative into the same domain to reduce the intra-class distance between the different domains of the same individual rigorously. In addition, an adaptive weight is applied to the loss of each part-representative vector generated by the part-relationship attention module (PRAM) which allows the consideration of different deviations for each component feature induced by pose and emotion variations.

$$L^i_{CAT} = \sum_{i=1}^{5}\lambda_i * \left[\frac{CS(x^a_i, x^p_i) + 1}{CS(x^a_i, x^n_i) + 1} - m\right]_+ \tag{15}$$

CS represents the cosine similarity, and m the conditional margin. x_i indicates the feature vector of each component, and a, p, n represents the anchor, positive and negative example, respectively. λ_i is the intersection over union of the extracted masks of the anchor and positive examples.

5. Deep Learning Methods

Deep learning is an effective approach for face recognition which has been shown to produce interesting results. Since its appearance in 2012 with AlexNet [84], which showed powerful learning capabilities, deep learning has become the go-to method for solving complex problems. Its ability to process large datasets makes it the first method to be applied today to biometric problems such as face recognition. This method contains many approaches to address face recognition problems, the most used to date being synthetic methods and VIS-NIR alignment methods. In this section we will list the different methods that help in IR face recognition, starting with synthetic methods, then discussing feature extraction and NIR-VIS alignment methods, and then presenting applications of these methods.

5.1. Synthesis Methods

This method performs face image synthesis from one spectrum to another. It links the domain divergence to the image preprocessing step. After the transformation into the new spectrum, it directly performs the face matching of the heterogeneous images. Reconstruction of the face image in the visual band allows for a more efficient extraction of facial features that will be used to classify and validate the images. By providing only IR images of faces as probes, Lai et Yanushkevich [85] explored the possibility of verifying and identifying faces in the visible range. They considered the following approach to achieve this goal. A generative adversarial network (GAN) was used to generate visible images from thermal images, and then face recognition techniques are applied to the computer-generated images. They started with a set of visual and thermal images and then normalized pairs of these images using facial landmarks. After normalization, they applied a GAN to learn and generate the correspondence between the thermal and visual images. Finally, to evaluate the face recognition performance, they injected the synthesized visible images into three different CNNs: inceptionV3 [86], Xception [87] and MobileNet [88].

Litvin et al. [89] proposed an accurate deep network architecture for the reconstruction of RGB facial images to thermal images for use in face recognition. As shown in Figure 6, RGB image generation is performed by applying a modified FusionNet [90] architecture. To optimize the results, they used a decoding block with a resized convolution instead of a transposed convolution. Besides that, a drop-out block was added between the bridge and the new decoding block, and a Randomized leaky ReLU (Randomized Leaky ReLU) [91] replaced the standard rectified linear units (ReLU) to decrease the overfitting effect. Litvin et al. [89] trained a face classifier to test the reconstructed RGB images and compared the results with a reconstructed dataset using FusionNet as well as the original RGB images.

To simplify the synthesis of heterogeneous faces, He et al. [92] presented a high-resolution heterogeneous face synthesis divided into three main parts as shown in Figure 7; a Gp pose correction network that estimates normalized shape information, a Gt texture inpainting network that learns to produce a pose invariant facial texture representation, and a fusion warping network that combines the results of the previous two parts. To supervise the visual quality and minimize intra-class variance, they used a multi-scale and a fine-grained discriminator respectively. To obtain high synthesis results, three types of losses were employed: a UV loss, an adversarial loss, and a pixel loss.

Wu et al. [93] used the Disentangled Variation Representation (DVR) for intermodal NIR-VIS face matching. To disentangle the NIR and VIS facial representations, they implemented a variational lower bound to estimate the space of the posterior variable and the latent variable. To facilitate the modeling of the compact and discriminant disentangled latent variable spaces for heterogeneous modalities, they offered a way to minimize the identity information for the same subject and the relaxed correlation alignment constraint between the variations of the NIR and VIS modalities. They used the LightCNN-9 and LightCNN-29 [94] models as backbone networks. These are pre-trained on the MS-Celeb-1M dataset [95].

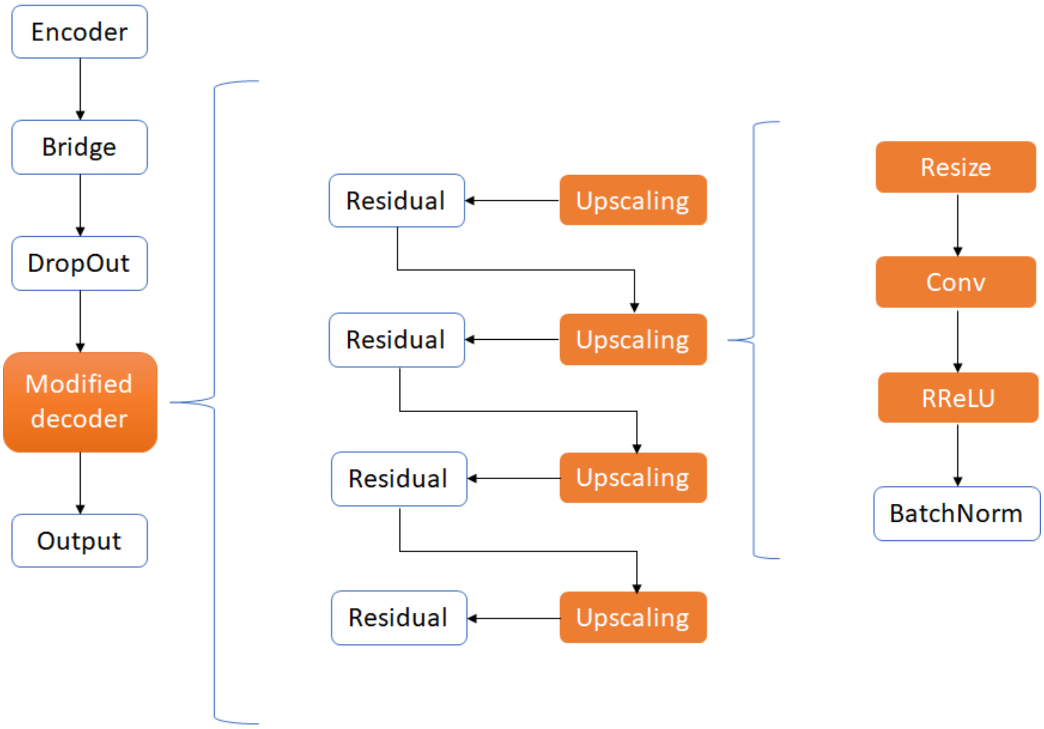

Figure 6. The proposed modified FusionNet based on the work of [89].

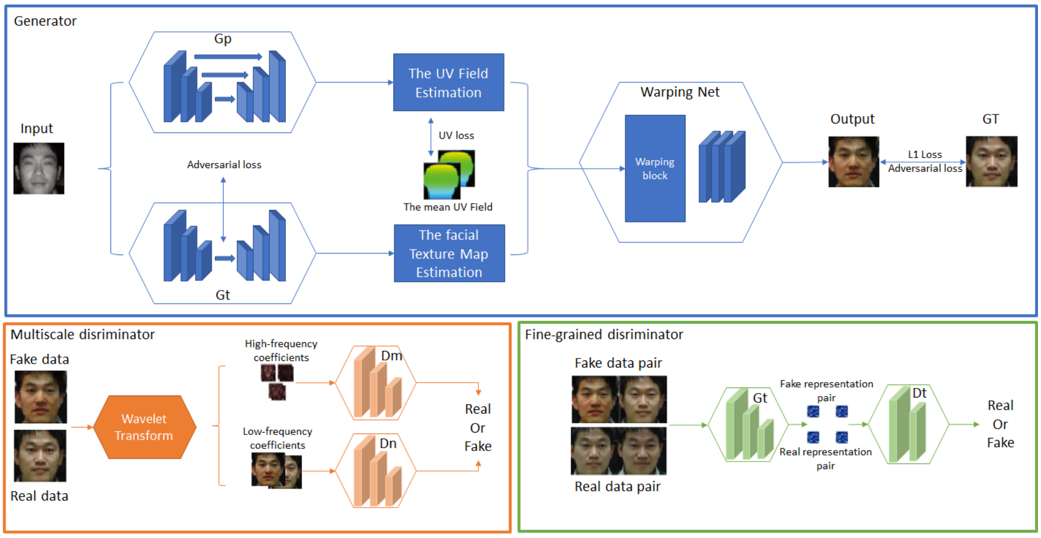

Figure 7. NIR-VIS face completion network based on [92].

One of the problems encountered in the identification or recognition of IR faces is the low resolution of IR spectral images. Existing approaches to infrared and visible face verification assume that infrared and visible images of faces have similar resolution. This is unlikely in real-world long-range surveillance systems because humans are far away from the cameras. Guei and Akhloufi [96] addressed this issue by providing a deep convolutional generative adversarial network (DCGAN) [97] for the enhancement of infrared facial images. The proposed algorithm synthesizes a super-resolution facial image from its lower resolution equivalent as shown in Figure 8.

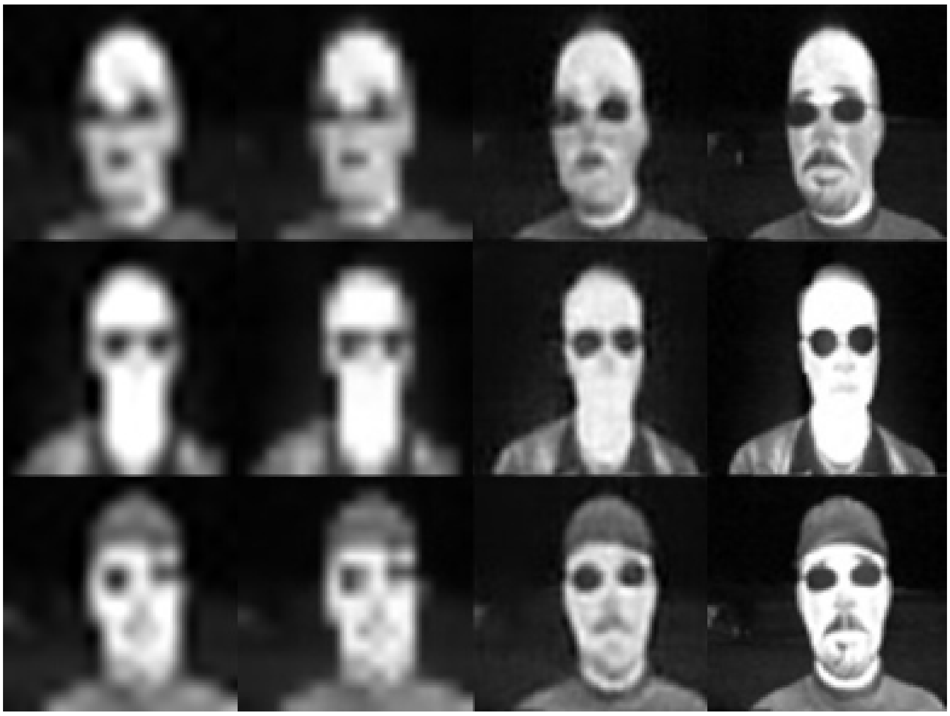

Figure 8. DeepSIRF 2.0 results for the thermal LWIR images, from left to right: low-resolution images; bicubic; DeepSIRF 2.0; high-resolution image [96].

To address the image resolution issue, Immidisetti et al. [98] introduced the task of thermal to visible face verification from low-resolution thermal images. The proposed task is difficult due to the significant domain disparity between the thermal and visual pictures and the low resolution of the thermal images. To address it, they suggested a hybrid network that combines axial-attention layers with an image conditional generative adversarial network (GAN). The generator generates images of visible faces which are then compared to a gallery of visible images using an available face matching algorithm.

Reflected light is created around the eyes in NIR face pictures of eyeglass-wearing subjects due to active NIR light sources and this is one of the primary performance degrading factors in NIR face recognition. To solve this problem, Kim et al. [99] proposed a Glasses2Non-glasses (G2NG) data augmentation. They adapted CycleGAN to implement synthetic oversampling and generated realistic facial images of subjects with and without glasses. They then combined the synthetic images with the database to build an augmented training database which improved the reflected light resistance of the NIR face recognition system.

In the work of Luo et al. [100], a new generative adversarial network was proposed for facial image translation in thermal to RGB visible light named ClawGan: Claw connection-based generative adversarial networks. Luo et al. [100] proposed a mismatch metric (MM) to measure the mapping relationship of paired images and used template matching to reduce the MM of the dataset. To form a new objective function, they added the synthesized loss and the generative reconstructed loss to the adversarial loss and the cycle-consistency loss. And to improve feature preservation, they replaced the U-Net network structure of the generator with a claw-connected network. The proposed ClawGAN image translation system preserves the thermal properties of the images while improving their quality. It also improves the recognizability of the images in both dark and bright light.

5.2. Feature Learning Methods

The method of learning features consists of locating faces by extracting facial structural features. It is first trained as a classifier and then it is used to differentiate between facial and non-facial regions. The idea is to go beyond the limits of the instinctive knowledge of faces. In the work of Wu et al. [101], they introduced a convolutional neural network (CNN) architecture for thermal face recognition as a new approach to automatically learn efficient features from raw data. The results of experiments on the RGB-D-T face dataset show that the proposed CNN architecture can achieve a higher recognition rate compared to traditional recognition methods such as LBP, HOG [102], and invariant moments.

The high-level features of deep convolutional neural networks trained on images of visual spectra are potentially domain-independent and can be used to code for faces detected in different image domains. Pereira et al. [103] presented a generic framework of domain-specific units for the recognition of heterogeneous faces using a deep neural network architecture with low-level features. With this approach, the learning of shallow feature detectors of each new image domain is possible.

In Peng et al. [104], a convolutional neural network (CNN) for NIR face recognition (specifically face identification) in non-cooperative-user applications is presented. The proposed NIRFaceNet is a modified GoogLeNet [105] but it has a more compact structure specifically designed for the NIR dataset of the Chinese Academy of Sciences Institute of Automation (CASIA) and it can achieve higher rates of identification with less training time and less processing time. The experimental results show that when image blur and noise are present, NIRFaceNet has an overall advantage over other approaches in the face recognition domain of NIR images.

To extract modality-invariant and identity-discriminative features, Hu et al. [106] proposed a Disentangled Spectrum Variations Network. This deep learning framework deals with the NIR-VIS disentangle spectrum variations matching problem. To do so, two strategies are presented: Stepwise Spectrum Orthogonal Decomposition (SSOD) and Spectrum adversarial Discriminative Feature Learning (SaDF). The first one consists of assigning the task of disentangling spectrum variations to several layers of the network to model the process of layer-by-layer removal of spectrum information in the network. The second is to learn the identity discrimination features which consists of an identity discrimination subnetwork (IDNet) and an auxiliary spectrum opposition subnetwork (ASANet). IDNet is composed of a *Gh* generator to generate an invariant spectrum feature and a *Du* discriminator to extract the identity discrimination feature. ASANet contains a *Gh* generator to remove modality-variant spectrum information with the help of a *Dm* discriminator as presented in Figure 9.

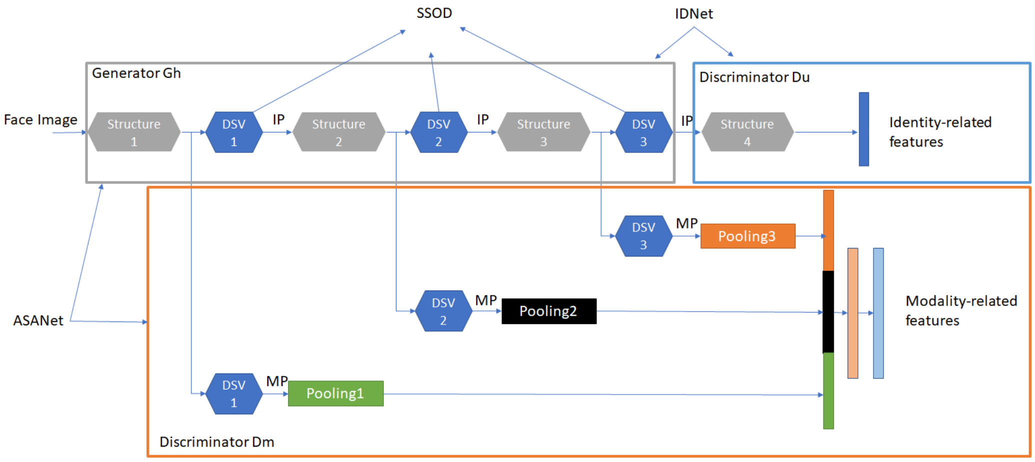

Figure 9. An illustration of the proposed DSVNs architecture based on [106].

Kim et al. [107] introduced an approach to solve the problem of applying a complex deep CNN architecture directly to NIR FR with limited size NIR face datasets. To improve the performance of FR NIR, a fine-tuning approach was used to pre-process the information of an FR RGB model from a deep CNN model using the pre-trained RGB parameters as the initial parameter for the NIR deep CNN model. The proposed approach achieved high performance with small public datasets and better generalization for various environments in a real-world FR scenario.

Before diving into more complex methods, some researchers choose to adapt existing methods from VIS FR to the IR FR. Shavandi and Paeen Afrakoti [108] studied the function of a sparse processing classification algorithm in thermal face recognition. This processing is applied directly to the input image to test the capacity of the sparse classifier to receive information directly from thermal images without using a feature extraction algorithm. The results showed that this algorithm successfully overcame challenges such as different facial states, images with and without glasses, images with noise present, thus outperforming the Eigenface and KNN algorithms.

By using a model trained on RGB images, Szankin et al. [69] studied the influence of thermal imagery resolution on the accuracy of Deep Learning-based face recognition. They used thermal images for the embedding phase which helped increase accuracy. A more efficient result was presented by using a new deep super-resolution model (SR) to enhance down-scaled images and increase accuracy by 6.5% on small data.

To perform face recognition on a thermal dataset, Mahouachi et Akhloufi. [109] developed a deep convolutional neural network architecture based on the FaceNet architecture and the MTCNN model. FaceNet was used as an embedder. The authors built an intermediate model by concatenating three outputs of FaceNet models. The model accepts three images as its input and creates three 128-D vector embeddings as its output. Its goal is then to freeze some levels of the first layers and retrain the last layers to perform the task. Later, Mahouachi et Akhloufi. [110] adapted the work of [109] on NIR images. The authors used raw data without image quality enhancement and chose not to use pre-trained weights on the RGB datasets and trained the model from scratch using multiple fine-tunings.

Jo et al. [111] conducted experiments on the application of deep learning on NIR face recognition. Two deep learning networks were trained (FaceNet and NIRFaceNet [104]) on five public datasets. This experiment showed that simple networks perform well on NIR face datasets as shown in Figure 10. They presented a data augmentation method to improve the recognition of users with glasses which helped to overcome this category's problems without constructing an additional training set.

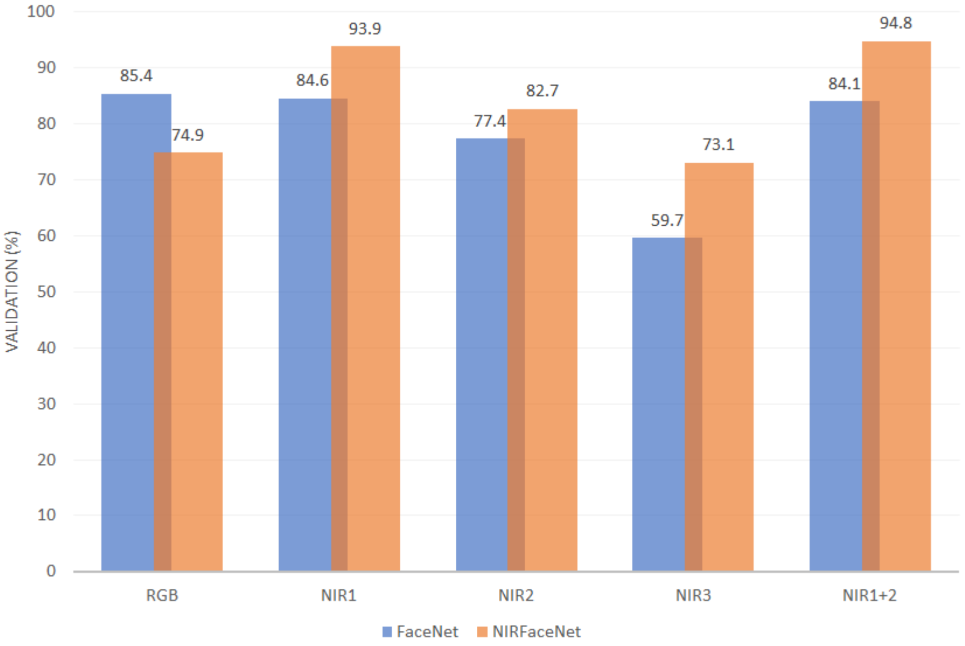

Figure 10. Comparisons between FaceNet and NIRFaceNet [111].

Also, Gavini et al. [112] proposed a method to improve thermal classifier accuracy by using transfer learning. They proposed a two-classifier technique for thermal face recognition in which the source classifier is taught first. The target classifier is then trained using the source classifier's information. The suggested approaches ($CNN_t^{DSD_{TL1}}$ and $CNN_t^{DSD_{TL2}}$), in which the fine-tuned weights of the sparsified source network are transferred, enhance the target classifier model's capacity (CNN_t). The results of these methods show an increase in the accuracy of thermal to visual face recognition.

Residual Compensation Networks (RCN) were introduced in [82]. The authors used a novel two-branch network architecture (RCN) to acquire separate features for different modalities in Heterogeneous Face Recognition (HFR). The RCN incorporates a residual compensation (RC) module and a modality discrepancy loss (MD loss) into traditional convolutional neural networks. The RC module reduces modal discrepancy by adding compensation to one of the modalities so that its representation can be close to the other modality. The MD loss alleviates modal discrepancy by minimizing the cosine distance between different modalities.

A two-step method is considered in the work of Guo et al. [62]. This paper uses public VIS data resources to train a deep network model which is referred to as the first model. As a next step, they used several near-infrared face images to retrain the obtained deep network model. After retraining is completed, they used the last deep network model as a feature extractor of near-infrared face images. Then they apply the cosine distance to calculate the score of both features for the test sample and training sample. Here, the score could be considered to be the correlation intensity between the test sample and training sample. They then used the weighted combination strategy to perform score fusion by applying an adaptive score fusion strategy and the nearest neighbor algorithm to conduct the final classification (See Figure 11).

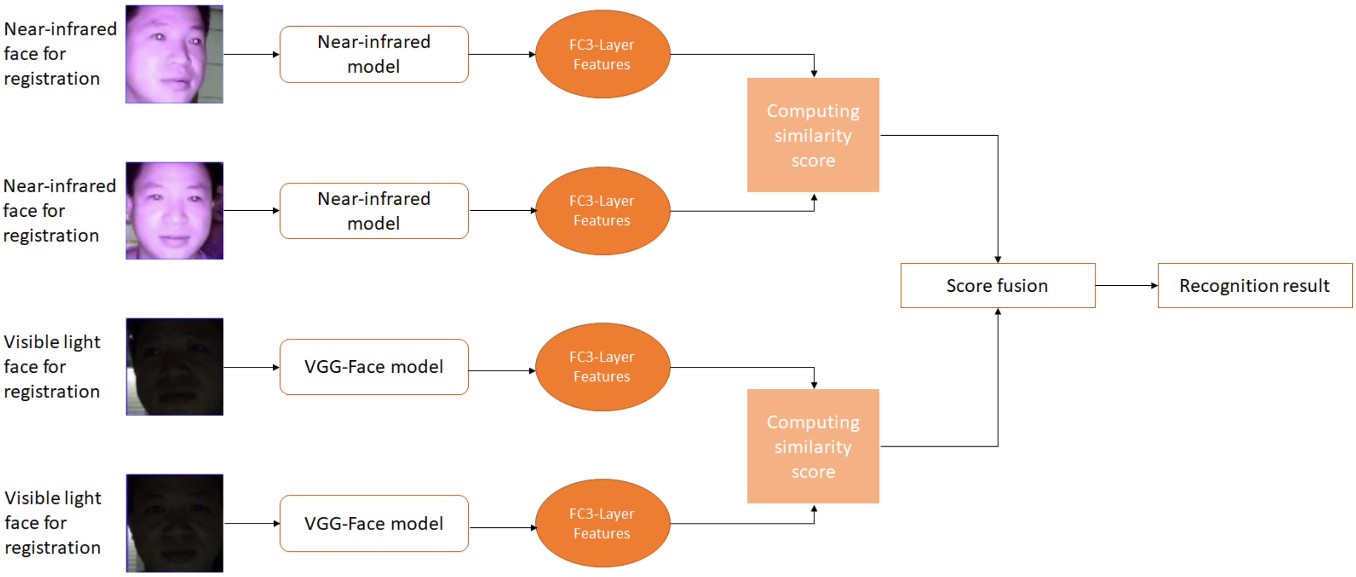
Near-infrared face for registration
Near-infrared model
FC3-Layer Features
Computing similarity score
Near-infrared face for registration
Near-infrared model
FC3-Layer Features
Score fusion
Recognition result
Visible light face for registration
VGG-Face model
FC3-Layer Features
Computing similarity score
Visible light face for registration
VGG-Face model
FC3-Layer Features

Figure 11. Score fusion process [62].

A novel CNN structure is proposed based on characteristics of thermal infrared faces [113]. Convoluted edges are taken as the initial features to refine and extract uncommon thermal infrared facial features for identification. This paper suggested a regional parallel structured CNN algorithm (RPSNet) to obtain multi-scale features based on edge information. The structure of the proposed network contains three main cascaded components as shown in Figure 12: initial edge feature extraction, multi-scale feature extraction, and feature vector classification. The initial edge feature extraction module is composed of a convolution layer and a cascaded maximum clustering layer. In order to generate multi-scale features, they design the convolutions with different kernels as three parallel channels. Finally, the fully connected layer and the softmax loss transform the convolutional feature maps into feature vectors.

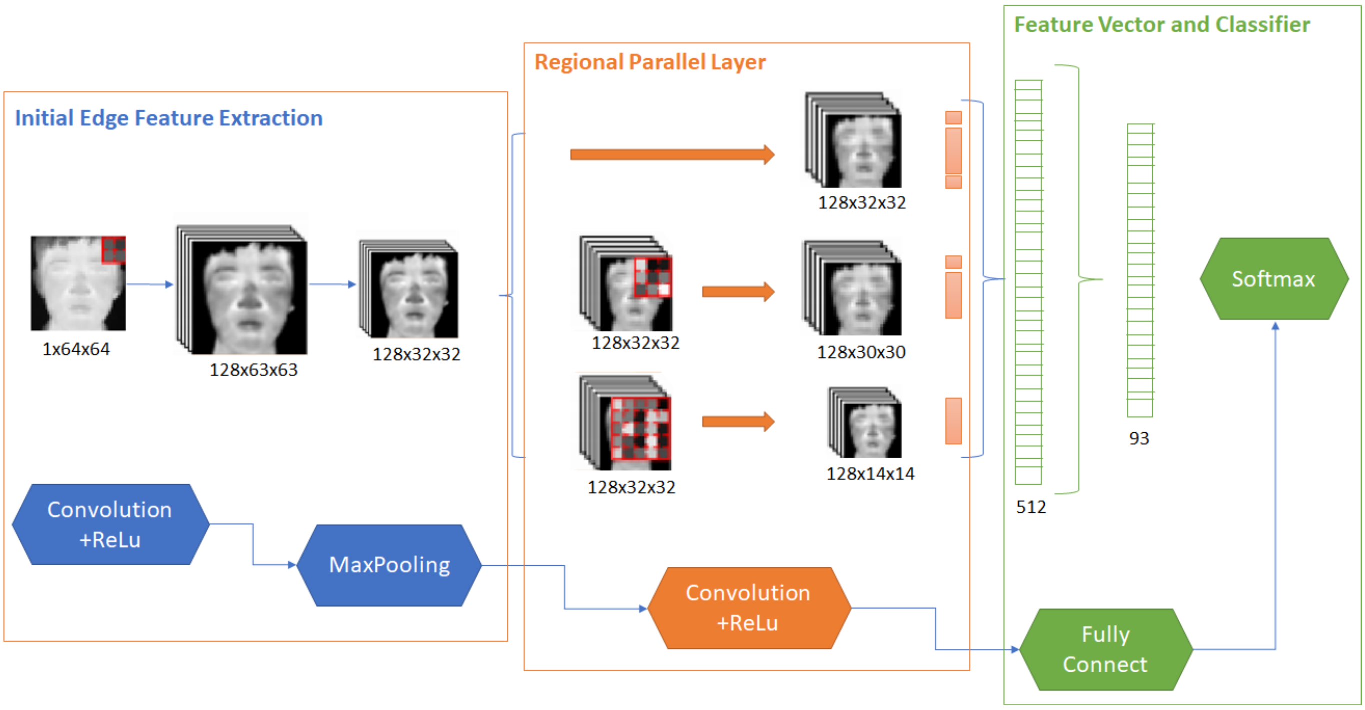

Figure 12. Regional parallel structure convolutional neural network based on [113].

A new Wasserstein convolutional neural network (W-CNN) is introduced in [114]. This approach learns the invariant features between near-infrared and visual images. The overall architecture is presented in Figure 13. The low-level layers of the W-CNN are trained with VIS images and the high-level layer is split into three parts: the NIR layer, the VIS layer, and the shared NIR-VIS layer. The first two layers aim to learn the specific features of each model and the NIR-VIS shared layer is built to learn a subspace of invariant features of the modality. The Wasserstein distance is introduced in the NIR-VIS shared layers to measure the similarity between distributions of heterogeneous characteristics. W-CNN learning is implemented to minimize the Wasserstein distance between the NIR and VIS distribution for invariant deep feature representations of heterogeneous face images. They imposed a correlation prior on the fully connected W-CNN layers to avoid the problem of over-fitting on small-scale heterogeneous face data by minimizing the size of the parameter space. This prerequisite is implemented by a lower rank constraint in an end-to-end array.

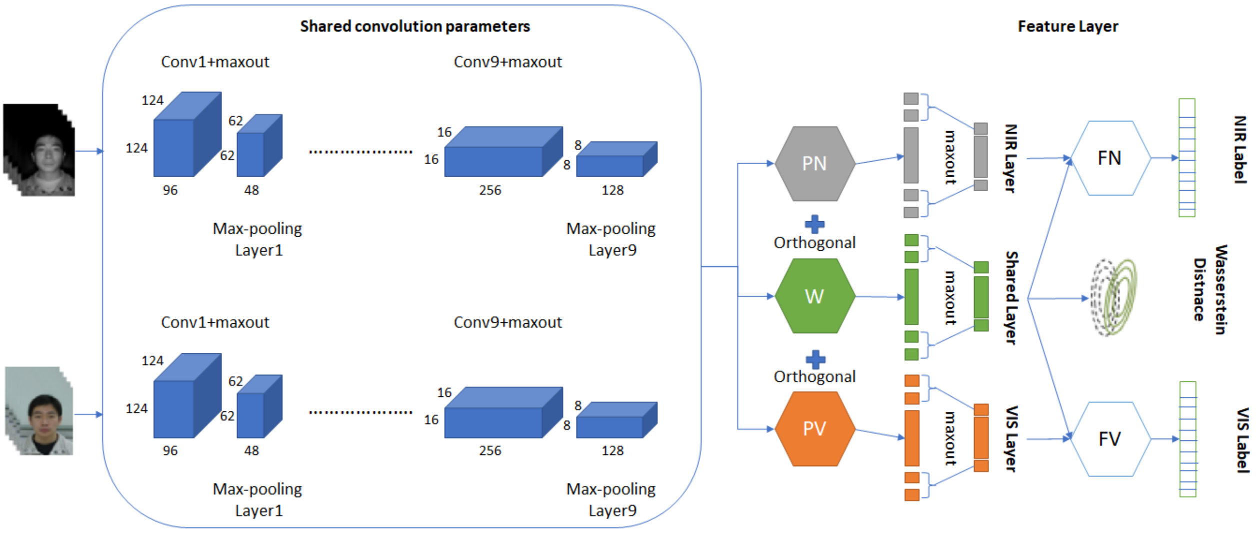

Figure 13. An illustration of the proposed Wasserstein CNN architecture in [114].

To extend the developments in deep learning for VIS face recognition to the NIR spectrum without having to reconsider the underlying deep models that can only see VIS faces, Lezama et al. [115] proposed an approach that includes two essential elements, cross-spectral hallucination and low-level embedding, to respectively optimize the input and output of a deep model using the VIS spectrum for cross-spectral face recognition. Cross-spectral hallucination generates VIS faces from NIR images using a deep learning approach. Low-level integration provides a low-level structure for the deep features of faces in the NIR and VIS spectrum.

No attempt was made to reduce the complexity of DCNN models for NIR face recognition until Kim et al. [99] suggested a fast NIR face recognition system based on the DCNN that is resistant to reflected light. They used Glasses2Non-glasses (G2NG) data augmentation to create synthetic face images of people with and without eyeglasses. Also, in this study, they produced LiNFNet by reducing the complexity of VGG-16 using depthwise separable convolutions and linear bottlenecks. The initial convolution layers are simpler functions for extracting output activations than the rest of the convolution layers. The output activations extracted from the initial convolution layers of VGG-16 for an NIR face image have similar patterns and structures of intensity values. From this analysis, they conclude that the activations contain redundant information. Thus, to implement the LiNFNet architecture, they reduced the number of filters in the first convolution layer of the network by half. Then, to extract the rich information for NIR FR from the input activation, they adapted linear bottlenecks to the last three convolution layers of VGG-16.

Due to the lack of datasets, heterogeneous FR techniques commonly rely on pre-trained features from a large-scale visual dataset, including generic face data. However, due to the texture mismatch with the visual domain, these pre-trained features result in worsened performance. Based on this reasoning, Cho et al. [116] presented the Relational Graph Module (RGM), a graph-structured module that collects global relational information as well as generic facial features. Since the relational information of each identity between intra-facial components is the same regardless of modality, understanding the relationship between features can facilitate cross-domain matching. Using GMR, relationship propagation reduces texture dependence while retaining the benefits of pre-trained features. In addition, the GMR identifies long-range connections by capturing global face geometries from locally linked convolutional features. The authors also proposed a node attention unit (NAU) which performs node-wise recalibration to focus on the most informative nodes emerging from the relationship-based propagation.

Kumar et al. [117] presented the Occluded Thermal Face Recognition using the Bag of CNN (BoCNN) architectural framework for recognizing occluded thermal faces. They began by examining the effectiveness of preprocessed models using transfer learning to find that they produce good results for thermal faces that are not occluded. Since occlusion reduces performance, they used other decision-level fusion techniques after transfer learning to improve the performance of the pre-trained models. Compared to a single CNN architecture, all fusion techniques used in the presented study produce better results. Several pre-trained serial DAG and CNN models, such as VGG-19, Resnet-50, Resnet-101, Inception-V3, and InceptionResnetV2, have been combined into the proposed BoCNN model as seen in Figure 14. The characteristics generated differ due to the differences in architectural design and depth of each CNN model. It has also been empirically determined that each network's misclassification is not mutually inclusive. These models were fused at the decision level utilizing two distinct types of fusion techniques. Various fusion techniques, namely majority voting, maximum score, and average score at level 1 and majority average and majority-maximum at level 2, were used to combine different networks formed after training.

Xu et al. [83] presented a part relationship attention module (PRAM) that extracts connections between components from a domain-independent face semantic mask to learn domain-independent features as well as fluctuations in pose and emotion. To express domain-invariant identification information, the relationships between facial components are crucial. PRAM involves four steps. According to a previously extracted mask, a face image is first cut into four parts: left eye, right eye, nose, and mouth. The light CNN-9 backbone receives the four partial images and the global image of the face for a total of five images. Representative features of each part are carefully extracted at this stage. The out-of-order pairwise combinations were retrieved and organized in a predefined order in the second phase to illustrate the relationship between two sections. All combinations are then sent to a common FC layer (L2) in the third phase, which ensures that the network learns the same functional connection between two representative features. The association between specific locations and a consistent standard is derived from this calculation. A learning weight is used in the final phase to reflect the strength of each relationship.

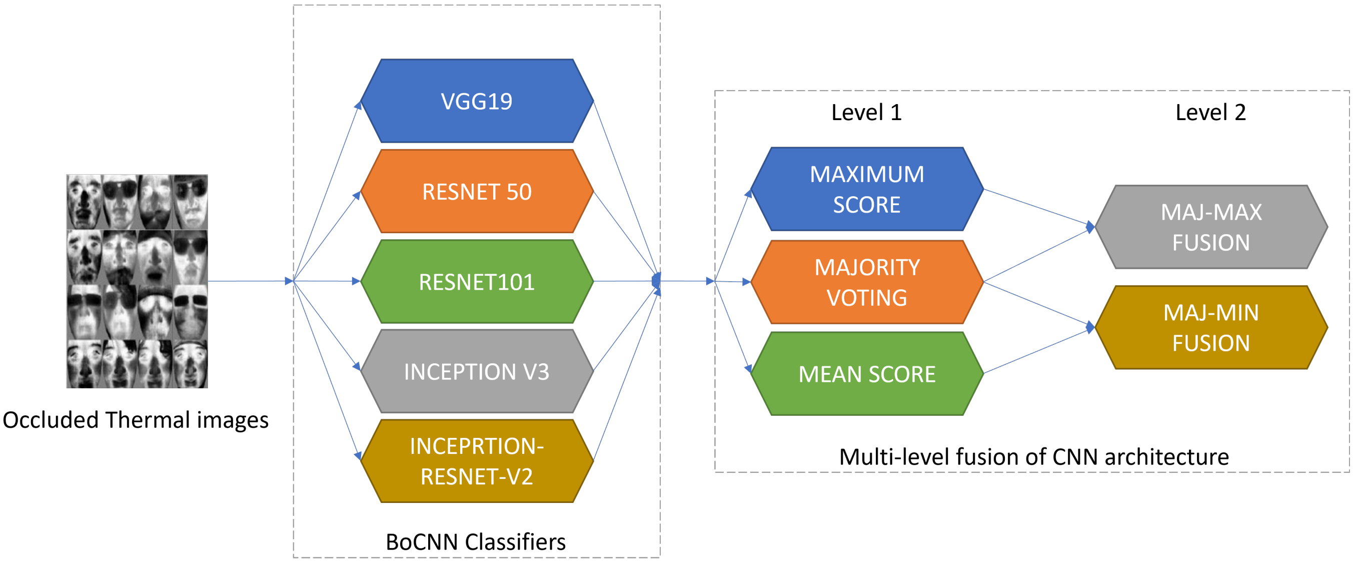

Figure 14. The BoCNN architecture proposed in [117].

5.3. NIR-VIS Alignment Methods

Near infrared-visual face recognition (NIR-VIS) is a task that involves matching face data from multiple modalities, and it has a wide range of applications in areas such as multimedia information retrieval and criminal investigations. However, due to significant intra-class variability and small NIR-VIS datasets, it remains a challenging task. Several methods are used for the alignment approach. Sarfraz et al. [118] presented the first attempt in using deep neural networks to bridge the modality gap in thermal-visible face recognition. The learned projection matrices capture the non-linear relationship and are able to bring the two closer to each other. In [119] a new invariant deep representation approach is presented by He et al. This method maps NIR and VIS images to a compact Euclidean space using a network that is composed of two layers. The low layers are trained on VIS data and the high layers contain two subspaces; a modality invariant identity information and a modality variant spectrum information. For optimization, an alternation of minimization is used at the training phase. Wu et al. [73] presented an image-image translation to enhance NIR face recognition. This method is divided in three sub-methods: face alignment by using the MTCNN network [120], NIR-VIS image translation using the CycleGan framework [121] to generate VIS images from NIR images, and an Inception-ResNet-v1 model to use as a face embedding based on FaceNet. This method shows an efficient way to transform NIR face images into VIS images by maintaining identity information needed for recognition. Deng et al. [81] propose a new heterogeneous face recognition modal invariant deep neural network. They first extract modal independent hidden factors for different modalities using a mutual component convolutional neural network layer instead of backpropagation to prevent overfitting. An MCA loss is then presented for modal invariant feature extraction of single images.

Considering a heterogeneous face recognition problem, the significant domain gap between the NIR and VIS modalities presents great challenges to accurate face recognition. To overcome the domain gap problem, Wang et al. [122] proposed a Parallel-Structure-based Transfer learning method (PST) (See Figure 15), which fully utilizes multi-scale feature map information. PST consists of two parallel streams of the network; a source stream (S-stream) with fixed parameters from being pre-trained on a large-scale VIS dataset and a transfer stream (T-stream) that absorbs multi-scale feature maps from S-stream and transfers the NIR and VIS face embeddings to a unique feature space. S-stream preserves the discriminative ability learned from the large-scale source dataset.

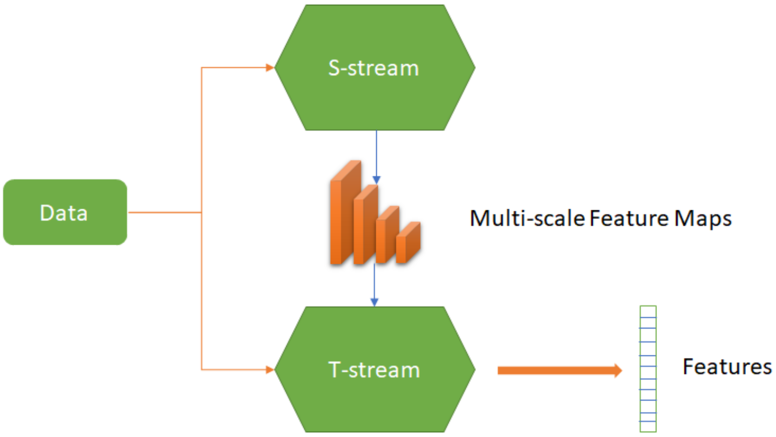

Figure 15. Parallel-Structure-Based Transfer learning method based on [122].

The use of unpaired VIS images to improve the NIR-VIS recognition accuracy is an ongoing issue. Liu et al. [123] presented a deep TransfeR NIR-VIS heterogeneous facE recognition neTwork (TRIVET) for NIR-VIS face recognition. First, a deep convolutional neural network (CNN) with ordinary measures was used to learn discriminative models to utilize large numbers of unpaired VIS face images. The ordinal activation function (Max-Feature-Map) was used to select discriminative features and to make the models robust and lighter. Second, to transfer these models to the NIR-VIS domain, they fine-tuned two types of NIR-VIS triplet losses. The triplet loss not only reduces intraclass NIR-VIS variations but also augments the number of positive training sample pairs. This makes fine-tuning deep models on a small dataset possible.

To address the misalignment problem, Zhao et al. [124] proposed a self-aligned generation architecture to semantically align data distributions between two modalities. To generate matched images with different domains, they used two encoders and two decoders. To train the network, a training method is promoted with the same latent code and a self-aligned block. While the self-aligned block works as a secondary rectifier of the unaligned attributes, the same latent code might virtually impact alignment performance. These methods ensure that images from two domains are aligned. In addition, they presented a multi-scale patch discriminator for the high quality of the generated aligned NIR-VIS images.

Hu et al. [125] presented a new method to solve the NIR-VIS matching problem called Dual Adversarial Disentanglement and Deep Representation Decorrelation (DADRD). Three major components of the DADRD model are designed to reduce the gap between NIR-VIS images: Cross-Modal Margin Loss (CmM), Dual Adversarial Disentanglement Variation (DADV) and Deep Representation Decorrelation (DRD). First, as shown in Figure 16, the CmM loss collects the intra-class and inter-class information from the data, and then uses a central variation element to reduce the modality difference. The mixed face representation (MFR) layer of the backbone is then divided into three sections: the identity-related layer, the modality-related layer, and the residual-related layer. The DADV is intended to decrease intra-class variations such as adversarial disentangled modality variations (ADMV) and adversarial disentangled residual variations (ADRRV). Specifically, ADMV and ADRRV use an adversarial method to remove spectral and residual variations such as lighting, posture, emotion, and occlusion. Finally, they apply DRD to the three deconstructed features to make them uncorrelated which more effectively separates information from the three components and improves feature representations.

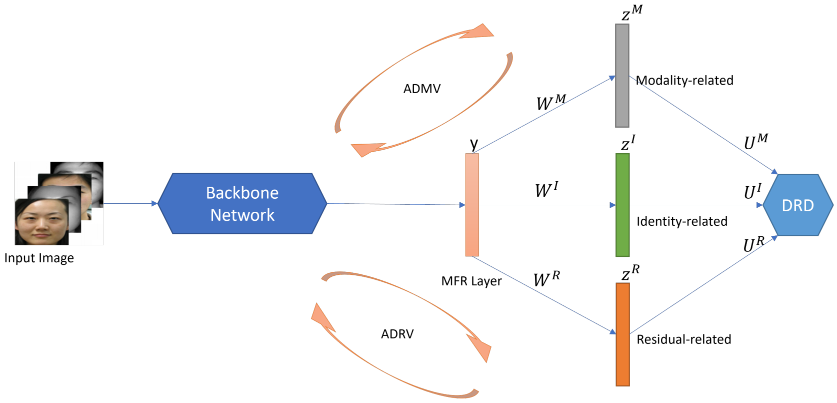

Figure 16. The flowchart of DADRD model proposed in [125].

To reduce the difficulty of learning cross-mode invariant features, Sun et al. [126] proposed a method of decomposing the cross-mode data gap by auxiliary modality (DGD) for HFR NIR-VIS. The authors used the brightness component (Y component) of the YCbCr color space of the VIS image to decompose the cross-mode data gap as an auxiliary modality. The huge gap between the NIR data and the VIS data is decomposed into two smaller gaps. This is because the brightness component retains the structural information of the VIS image and it is similar to the color information of the NIR modality which reduces the difficulty of network learning. Then, during the learning process, they designed the intermodal gap decomposition loss and intramodal gap loss to guide intermodal knowledge exchange and back-propagation optimization.

Recently, Cheema et al. [127] proposed a unified end-to-end cross-modality discriminator network (CMDN) for HFR. This work presents a cross-modality discriminator network and unitary class loss for heterogeneous face recognition. In order to learn deep relationships between features for cross-domain face matching, the proposed network uses a deep relational discrimination module. At the same time, it is used to extract modality-independent embedding vectors for face images. The unit class loss aids in the parameter optimization of the CMDN network which shows high stability and accuracy. The proposed loss can learn invariant identity features from unaligned facial images. This network can be used not only to extract embedding vectors from faces but also to perform HFR classification and to create a fusion of embedding vectors and classification probabilities.

5.4. Applications

Infrared Face recognition has been tested in real-world situations. Menon et al. [128] proposed a solution to identify drunk drivers using images captured with the thermal spectrum. Two steps are necessary. First, face identification is performed to detect faces using a CNN. The detected face is classified into one of two classes, drunk or sober, using a Gaussian Mixture Model along with the Fisher Linear Discriminant for dimensionality reduction. To do so, a set of selected points on the same faces with different levels of alcohol show the temperature distribution of the face allowing the classification of the images. Kamath et al. [129] presented TERNet, a powerful emotion detection system based on thermal pictures. To achieve this, a modified convolutional neural network with strong generalization characteristics was used. The architecture uses features from the VGG-Face CNN model that were learned using transfer learning and fine-tuned with thermal expression face data from the TUFTS face database. Mohamed et al. [130] presented a multi-spectral face anti-spoofing method working with both visible (VIS) and near-infrared (NIR) spectra imaging. A novel solution based on active near-infrared images is proposed for face spoofing detection. Unlike most existing spoofing detection techniques, they used the NIR spectrum and they analyzed differential images. To detect the existence of spoofing media, they calculated the context consistency between face and non-face areas. Extreme circumstances, such as cropped fake media, necessitate the use of lighting texture.

Since the newly emerged difficulties during the COVID-19 pandemic, individuals are required to wear facial masks during this period to prevent the virus from spreading. From the perspectives of training data and training techniques, Du et al. [131] tackled the problem of NIR-VIS masked face recognition with the aid of semi-Siamese networks. They presented a unique heterogeneous training technique to optimize the mutual information provided by the face representation of two domains. Furthermore, to synthesize a masked face from an existing NIR image, a method based on 3D face reconstruction is used. Using these techniques, the approach generates a domain invariant face representation that is robust to mask occlusion. Tests on three NIR-VIS face datasets indicate the effectiveness of the method and its ability to be generalized to other datasets. It achieves a validation rate of 98.58%, 83.0% and 70.6% respectively on the CASIA NIR VIS dataset, the Oulu-CASIA NIR-VIS dataset and the BUAA-VisNir dataset.

6. Discussion

Infrared facial identification is able to overcome the limitations of visual face recognition such as changes in light. Infrared (IR) imaging in FR systems has attracted increasing interest due to the consistant quality of the images acquired under various lighting conditions. In this section, we summarize and discuss all the findings based on the reviewed papers. For that purpose we present Tables 2–5 to summarize the results.

Infrared recognition of faces suffers from several problems. One of the major problems is the lack of datasets containing IR images. The size of current IR face datasets is only one tenth the size of the well-known CASIA WebFace dataset for RGB FR. For that reason, a lot of work was oriented towards synthesis methods to overcome this problem. The idea is to generate visible images from infrared images to then perform face recognition using methods tested and trained over RGB images. We summarize the results obtained by these approaches in Table 2.

Table 2. Comparison of synthesis approaches results.

References	Methods	Metrics	Datasets
Lai and Yanushkevich [85]	CycleGAN InceptionV3 Xception MobileNet	95.35% (Rank-1 acc)	Carl dataset [70]
Litvin et al. [89]	FusionNet+RReLu VGG classifier	97.52% (Rank-1 acc)	RGB-D-T [61]
He et al. [92]	CFC (pose correction +texture inpainting +fusion wrapping)	99.21% (Rank-1 acc) 99.70% (Rank-1 acc) 99.90%\(Rank-1 acc)	CASIA NIR VIS 2.0 [53] BUAA-Vis-Nir [66] Oulu-Casia [35]
Wu et al. [93]	DVR (LightCNN-9, LightCNN-29)	99.10% 99.70% (Rank-1 acc) 99.30% 100.00% (Rank-1 acc) 97.90% 99.20% (Rank-1 acc)	CASIA NIR VIS 2.0 [53] Oulu-CASIA [35] BUAA-VIS-NIR [66]
Guei and Akhloufi [96]	DCGAN (DeepSIRF2.0)	243.21 (MSE) 140.16 (MSE) 140.16 (MSE)	Terravic Facial IR [71] CBSR NIR [53] CASIA NIR VIS 2.0 [53]
Immidisetti et al. [98]	Axial-attention layers C-GAN	94.40% (AUC)	ARL-VTF dataset [65]
Kim et al. [99]	Glasses2Non-glasses (G2NG) data augmentation CycleGAN	94.60% (VR@FAR + 0.1%)	LFW [46]
Luo et al. [100]	Claw-GAN	95.70% (AUC)	IRIS dataset [56]

Table 3. Comparison of feature extraction approaches results.

References	Methods	Metrics	Datasets
Zhan Wu et al. [101]	CNN	98,00% (acc)	RGB-D-T [61]
Pereira et al. [103]	DCNN (Inception Resenet v2 + adapting bias and kernels)	90.10% (Rank-1 acc) 92.20% (Rank-1 acc) 50.90% (Rank-1 acc)	CASIA NIR-VIS2.0 [53] NIVL NIR VIS [67] Pola Thermal [68]
Peng et al. [104]	Modified GoogleLeNet (NIRFaceNet)	98.28% (acc)	CASIA NIR [53]
Hu and Hu [106]	Stepwise spectrum orthogonal decomposition (SSOD), spectrum adversarial discriminative feature learning(SaDF) (IDNet, ASANet)	99.00% (Rank-1 acc) 100.00% (Rank-1 acc)	CASIA NIR VIS 2.0 [53] Oulu-CASIA [35]
Kim et al. [107]	Fine tuning pre-trained CNN models for RGB FR (FaceNet)	94.47% (VR@FAR = 0.7%)	PolyU-NIRFD [54]
Shavandi andAfrakoti [108]	Sparse processing classification (minimizing normed zero-norm, orthogonal matching pursuit)	96.50% (acc without any noise)	USTC NVIN [55] CBSR NIR [53]
Szankin et al. [69]	DNN (FaceNet) Face enhancement	99.33% (acc) 81.87% (acc)	SC3000DB [69] IRIS [56]
Mahouachi et Akhloufi [109]	FaceNet MTCNN Fine tuning	88.81% (VR@FAR = 50.66%)	USTC-NVIE [55]
Mahouachi et Akhloufi [110]	FaceNet MTCNN Fine tuning	96.68% (VR@FAR = 0.001%) 94.57% (VR@FAR = 49.01%)	CASIA NIR VIS 2.0 [53] USTC-NVIE [55]
Jo and Kim [111]	FaceNet NIRFaceNet Data augmentation	94.80% (VR@FAR = 0.1% without augmentation) 96.40% (VR@FAR + 0.1% with augmentation)	CASIA NIR-VIS 2.0 [53] +PolyU-NIRFD [54] +ND-NIVL [67]
Gavini et al. [112]	Transfer learning	94.32% (acc) 90.33% (acc)	RGB-D-T [61] UND-X1 [63]
Deng et al. [82]	Residual Compensation Convolutional Neural Network, Modality Descripency loss	99.32% (Rank-1 acc) 99.44% (Rank-1 acc)	CASIA NIR VIS 2.0 [53] CUHK NIR VIS [132]
Guo et al. [62]	DNN Cosine distance Adaptive score fusion	99.56% (acc weak light), 95.31% (acc strong light) 99.89% (acc weak light), 93.98% (acc strong light)	Sun Win [62] HIT LAB2 [57]
Wang and Bai [113]	RPSNet (edge feature extraction, multi-scale feature extraction, feature vector classification)	95.97% (acc)	Private dataset
He et al. [114]	W-CNN Low rank correlation constraint	98.70% (Rank-1 acc) 98.00% (Rank-1 acc) 97.40%\(Rank-1 acc)	CASIA NIR VIS 2.0 [53] Oulu [35] BUAA NIR VIS [66]
Lezama et al. [115]	Deep Cross-spectral Hallucination Low Rank Embedding	96.41% (acc)	CASIA NIR VIS 2.0 [53]
Kim et al. [99]	Lighten DCNN	94.60% (VR@FAR + 0.1%)	LFW [46]
Cho et al. [116]	Relational Graph Module	95.97% (VR@FAR = 0.1%) 99.22% (VR@FAR = 1%)	CASIA NIR VIS 2.0 [53] BUAA NIR VIS [66]
Kumar et al. [117]	Bag of CNN(BoCNN) VGG-19, Resnet-50, Resnet-101, Inception-V3, InceptionResnetV2	99.20% (mean score acc)	IIIT Delhi occluded thermal face dataset [72]
Xu et al. [83]	Part relationship attention module (PRAM) lightCNN-9	97.94% (VR@FAR = 0.1%) 98.44% (VR@FAR = 1%)	CASIA NIR VIS 2.0 [53] BUAA NIR VIS [66]

Another solution adapted for the lack of large dataset is a simple end-to-end IR feature extraction using transfer learning. This method uses pretrained weights on RGB images as input and then fine-tunes the model using an IR dataset to train an IR face recognition architecture. We summarize these approaches results in Table 3.

Table 4. Comparison of NIR VIS alignement approaches results.

References	Methods	Metrics	Datasets
Sarfraz and Stiefelhagen [118]	Feed Forward DNN Non-linear mapping	83.73% (Rank-1 acc)	UND X1 [63]
He et al. [119]	DNN Orthogonal subspace embedding	95.82% (VF@FAR = 0.1%)	CASIA NIR VIS 2.0 [53]
Wu et al. [73]	MTCNN CycleGAN	99.80% (acc) 99.60% (acc on Lab1), 90.70% (acc on Lab2)	INF [73] CSIST [57]
Deng et al. [81]	Mutual Component Convolutional Neural Network, MCA loss	99.22% (Rank-1 acc) 99.44% (Rank-1 acc)	CASIA NIR-VIS2.0 [53] CUHK NIR VIS [132]
Wang et al. [122]	Transfer Learning Multi-Scalefeature mapping	99.96% (Rank-1 acc)	CASIA NIR VIS 2.0 [53]
Xiaoxiang Liu et al. [123]	DNN Max-Feature-Map Fine-tuning Triplet loss	95.74% (Rank-1 acc) 91.03% (VR@FAR = 0.1%)	CASIA NIR VIS 2.0 [53]
Zhao et al. [124]	Self-aligned generation architecture Multi-scale patch discriminator	99.60% (VR@FAR = 0.1%) 93.20% (VR@FAR = 0.1%) 97.30% (VR@FAR = 0.1%)	CASIA NIR VIS 2.0 [53] Oulu CASIA [35] BUAA NIR VIS [66]
Hu et al. [125]	Dual Adversarial Disentanglement and Deep Representation Decorrelation	97.60% (VR@FAR = 0.1%) 92.90% (VR@FAR = 0.1%) 99.30% (VR@FAR = 0.1%)	CASIA NIR VIS 2.0 [53] Oulu CASIA [35] BUAA NIR VIS [66]
Sun et al. [126]	Dual Adversarial DGD	99.80% (VR@FAR = 1%) 85.30% (VR@FAR = 1%)	CASIA NIR VIS 2.0 [53] Oulu CASIA [35]
Cheema et al. [127]	End-to-end cross-modality discrimination network for HFR Unit-Class Loss	95.21% (Rank-1 acc) 98.50% ((Rank-1 acc) 99.70% (Rank-1 acc) 99.50% (Rank-1 acc)	TUFTS [74] UND-X1 [63] USTC-NVIE [55] CASIA NIR VIS 2.0 [53]

Table 5. Comparison of results on real-word application.

References	Methods	Metrics	Datasets
Menon et al. [128]	CNN Gaussian mixture model Fisher Linear Discriminant	97.00% (acc)	Private Dataset
Kamath et al. [129]	CNN Transfer Learning	96.20% (acc)	TUFTs dataset [74]
Mohamed et al. [130]	CNN	96.78% (acc)	Msspoof Dataset [133]
Du et al. [131]	Heterogeneous semi-Siamese method 3D face reconstruction	98.58% (VR@FAR = 0.1%) 83.0 % (VR@FAR = 0.1%) 70.6% (VR@FAR = 0.1%)	CASIA NIR VIS 2.0 [53] Oulu CASIA [35] BUAA NIR VIS [66]

One additional problem for the infrared spectrum is the occlusion due to the opacity of the glasses. Glasses lead to the occlusion of a large part of the face, resulting in the loss of important discriminant information. People with identical facial parameters can have different heat signatures. MWIR and LWIR images are sensitive to the ambient temperature, as well as to the emotional, physical, and health status of people. The consumption of alcohol alters the thermal signature of people, which can lead to performance degradation as shown in the work of Mahouachi et al. [109]. Yet, Menon et al. [128] used this property to identify drunk drivers using thermal images with a 97.00% accuracy as shown in Table 5.

For the synthesis methods, the best accuracy is obtained by Wu et al. [93]. Tested on CASIA NIR-VIS2.0, Disentangled Variational Representation using LightCNN-9 achieved 99.1% on rank 1 accuracy. When the backbone is replaced by LightCNN-29, DVR gains a further 0.6% on rank 1 accuracy. For the Oulu-CASIA NIR-VIS and BUAA-VIS-NIR datasets, while the quantity of samples in the training set is not large enough, they achieved 99.30% and 97.90% on LightCNN-9 and a gain of 0.7% and 1.3% on LightCNN-29 respectively.

A highly structured method that rivals the results of [93] is the Cross-spectral Face Completion used in [92]. The improved performance of CFC benefits in part from the use of Light CNN. They obtained 99.21%, 99.70% and 99.90% respectively on CASIA NIR-VIS2.0, BUAA-VIS-NIR and Oulu-CASIA NIR-VIS.

Although CycleGAN was developed for unpaired or unsupervised image synthesis, large face variations like pose and expression make CycleGAN fail to capture all the differences between the NIR and VIS domains. For that reason, Lai and Yanushkevich [85] only obtained a performance of 95.36% on the Carl dataset.

Most of the best performing techniques for feature extraction use small or private datasets. For this reason, we will only discuss performance on larger datasets. We find that Kim et al. [107], Deng et al. [82], and Hu and Hu [106] have the best rank 1 accuracy surpassing 99.0% accuracy. By fine-tuning the pre-trained RGB CNN model, Kim et al. [107] achieved 99.70% accuracy on the PolyU NIR face dataset. Gavini et al. [112] used transfer learning and achieved 94.32% on RGB-D-T and 90.33% on the UND-X1 dataset. For NIR-VIS alignment, transfer learning outperforms all other methods with accuracies of 99.96% in the work of Wang et al. [122] and 99.80% in the work of Wu et al. [73].

Briefly, this survey revealed that it is challenging to design a robust and reliable face representation that is based on both local and global features (hybrid method) in the NIR domain which is crucial for accurate FR. We expect that hybrid methods could offer better performance than methods based on single-type features. In Table 4, good recognition rates were reported using hybrid methods for IR FR.

A notable limitation of most of the studies reviewed is that while each of the previous methods was accurate in the presence of some challenges, their accuracy diminished in the presence of other challenges. For example, an invariant feature in facial expressions or eyeglasses works well as long as there is no misalignment or noise. The main reason is that most of the related work in the field of IR has focused on the problem of lighting. However, far too little attention has been paid to the other problems, such as noise, misalignment, and occlusion, which can occur in IR FR systems. For future work, we propose to incorporate image reconstruction and high-resolution methods into IR FR.

Since the different models were tested on different datasets with different augmentations, the comparison between their performances is not simple. Still, the analysis helps point out some of the important characteristics of a robust deep learning architecture for IR facial detection/recognition.

7. Conclusions

In this study, we have presented various deep IR face recognition and detection models that share the goal of improving the facial detection results. It is found that deep learning has achieved a degree of accuracy that facilitates deployment as a powerful tool that can be considered for many security applications. Research reveals that deep learning approaches have achieved high success rate using IR images for facial detection and recognition. Deep convolutional neural networks (CNNs) form the basis of the proposed techniques. For facial detection and recognition in IR, potential future studies could improve convolutional architectures. The performance could be improved by introducing new architectures and by analyzing current architectures. Both local and global features are used for recognition. Methods based on local features were found to be more efficient than global features. When visible and IR imaging is available, fusion and DNN-based methods are used to find the common feature space. A comparative analysis of the best performing architectures on the available datasets using similar metrics can help in their evaluation. Finally, one of the current limitations in this field is data accessibility. With more datasets we believe better results can be obtained. The results published to date show that the research in IR FR is still in its early stage but that it has the potential to improve significantly.

Author Contributions: Conceptualization, M.A.A. and D.M.; methodology, D.M. and M.A.A.; validation, D.M. and M.A.A.; formal analysis, D.M. and M.A.A.; writing—original draft preparation, D.M.; writing—review and editing, M.A.A.; funding acquisition, M.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was enabled in part by support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number RGPIN-2018-06233.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Jean-François Laplante for proofreading this paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

FR	Face Recognition
HO	Head Orientation
FE	Facial Expression
NIR	Near Infrared
VIS	Visible
MWIR	Middle Wavelength Infrared
LWIR	Long Wavelength Infrared
VIS FR	Visible Face Recognition
IR FR	Infrared Face Recognition
NIR FR	Near Infrared Face Recognition

References

1. Turk, M.; Pentland, A. Face recognition using eigenfaces. In Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Maui, HI, USA, 3–6 June 1991; pp. 586–591.
2. Sun, Y.; Liang, D.; Wang, X.; Tang, X. DeepID3: Face Recognition with Very Deep Neural Networks. *arXiv* **2015**, arXiv:1502.00873.
3. AbdAlmageed, W.; Wu, Y.; Rawls, S.; Harel, S.; Hassner, T.; Masi, I.; Choi, J.; Lekust, J.; Kim, J.; Natarajan, P.; et al. Face recognition using deep multi-pose representations. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–9.
4. Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; Song, L. SphereFace: Deep Hypersphere Embedding for Face Recognition. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6738–6746.
5. Mittal, S.; Agarwal, S.; Nigam, M.J. Real Time Multiple Face Recognition: A Deep Learning Approach. In *Proceedings of the 2018 International Conference on Digital Medicine and Image Processing*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 70–76.
6. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 815–823.
7. Ghiass, R.S.; Arandjelovic, O.; Bendada, H.; Maldague, X. Infrared face recognition: A literature review. *arXiv* **2013**, arXiv:1306.1603.
8. Zhao, W.; Chellappa, R.; Phillips, P.J.; Rosenfeld, A. Face Recognition: A Literature Survey. *ACM Comput. Surv.* **2003**, *35*, 399–458. [CrossRef]
9. Akhloufi, M.; Bendada, A.; Batsale, J.C. State of the art in infrared face recognition. *Quant. Infrared Thermogr. J.* **2008**, *5*, 3–26. [CrossRef]
10. Farokhi, S.; Flusser, J.; Ullah Sheikh, U. Near infrared face recognition: A literature survey. *Comput. Sci. Rev.* **2016**, *21*, 1–17. [CrossRef]
11. Ouyang, S.; Hospedales, T.; Song, Y.Z.; Li, X.; Loy, C.C.; Wang, X. A survey on heterogeneous face recognition: Sketch, infra-red, 3D and low-resolution. *Image Vis. Comput.* **2016**, *56*, 28–48. [CrossRef]
12. Jin, X.; Jiang, Q.; Yao, S.; Zhou, D.; Nie, R.; Hai, J.; He, K. A survey of infrared and visual image fusion methods. *Infrared Phys. Technol.* **2017**, *85*, 478–501. . [CrossRef]
13. Dey, T. A survey on different fusion techniques of visual and thermal images for human face recognition. *Int. J. Electron. Commun. Comput. Eng.* **2013**, *4*, 10–15.

14. Kakkirala, K.R.; Chalamala, S.R.; Jami, S.K. Thermal Infrared Face Recognition: A Review. In Proceedings of the 2017 UKSim-AMSS 19th International Conference on Computer Modelling & Simulation (UKSim), Cambridge, UK, 5–7 April 2017; pp. 55–60. [CrossRef]
15. Turk, M.; Pentland, A. Eigenfaces for Recognition. *J. Cogn. Neurosci.* **1991**, *3*, 71–86. [CrossRef] [PubMed]
16. Etemad, K.; Chellappa, R. Discriminant analysis for recognition of human face images. *J. Opt. Soc. Am. A* **1997**, *14*, 1724–1733. [CrossRef]
17. Liu, C.; Wechsler, H. Comparative assessment of independent component analysis (ICA) for face recognition. In Proceedings of the International Conference on Audio and Video Based Biometric Person Authentication, Washington, DC, USA, 22–23 March 1999; pp. 22–24.
18. Jonsson, K.; Matas, J.; Kittler, J.; Li, Y. Learning support vectors for face verification and recognition. In Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), Grenoble, France, 28–30 March 2000; pp. 208–213. [CrossRef]
19. Heo, J.; Kong, S.; Abidi, B.; Abidi, M. Fusion of Visual and Thermal Signatures with Eyeglass Removal for Robust Face Recognition. In Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop, Washington, DC, USA, 27 June–2 July 2004; p. 122. [CrossRef]
20. Chen, X.; Jing, Z.; Xiao, G. Fuzzy Fusion for Face Recognition. In *Proceedings of the Fuzzy Systems and Knowledge Discovery*; Wang, L., Jin, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2005; pp. 672–675. [CrossRef]
21. Ralescu, D.; Adams, G. The fuzzy integral. *J. Math. Anal. Appl.* **1980**, *75*, 562–570. [CrossRef]
22. Akhloufi, M.A.; Bendada, A. Fusion of active and passive infrared images for face recognition. In *Proceedings of the Thermosense: Thermal Infrared Applications XXXV*; Stockton, G.R., Colbert, F.P., Eds.; International Society for Optics and Photonics, SPIE: Baltimore, MD, USA, 2013; Volume 8705, pp. 84–93. [CrossRef]
23. Bebis, G.; Gyaourova, A.; Singh, S.; Pavlidis, I. Face recognition by fusing thermal infrared and visible imagery. *Image Vis. Comput.* **2006**, *24*, 727–742. [CrossRef]
24. Kong, S.G.; Heo, J.; Boughorbel, F.; Zheng, Y.; Abidi, B.R.; Koschan, A.; Yi, M.; Abidi, M.A. Multiscale Fusion of Visible and Thermal IR Images for Illumination-Invariant Face Recognition. *Int. J. Comput. Vis.* **2007**, *71*, 215–233. [CrossRef]
25. Heo, J.; Abidi, B.; Paik, J.; Abidi, M. Face recognition: Evaluation report for FaceIt identification and surveillance. In Proceedings of the SPIE 5132, Sixth International Conference on Quality Control by Artificial Vision, Gatlinburg, TE, USA, 1 May 2003. [CrossRef]
26. Akhloufi, M.A.; Bendada, A.; Batsale, J.C. Multispectral face recognition using non linear dimensionality reduction. In *Proceedings of the Visual Information Processing XVIII*; Rahman, Z.U., Reichenbach, S.E., Neifeld, M.A., Eds.; International Society for Optics and Photonics, SPIE: Orlando, FL, USA, 2009; Volume 7341, pp. 152–161. [CrossRef]
27. Brahnam, S.; Jain, L.C.; Nanni, L.; Lumini, A. *Local Binary Patterns: New Variants and Applications*; Springer: Berlin/Heidelberg, Germany, 2014. [CrossRef]
28. Li, S.Z.; Chu, R.; Liao, S.; Zhang, L. Illumination Invariant Face Recognition Using Near-Infrared Images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 627–639. [CrossRef]
29. Belhumeur, P.; Hespanha, J.; Kriegman, D. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 711–720. [CrossRef]
30. Gandhi, R. Boosting Algorithms: AdaBoost, Gradient Boosting and XGBoost. 2018. Available online: https://hackernoon.com/boosting-algorithms-adaboost-gradient-boosting-and-xgboost-f74991cad38c (accessed on 1 February 2023).
31. Méndez, H.; Martín, C.S.; Kittler, J.; Plasencia, Y.; García-Reyes, E. Face Recognition with LWIR Imagery Using Local Binary Patterns. In *Proceedings of the Advances in Biometrics*; Tistarelli, M., Nixon, M.S., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 327–336. [CrossRef]
32. Akhloufi, M.A.; Bendada, A. Infrared face recognition using texture descriptors. In *Proceedings of the Thermosense XXXII*; Dinwiddie, R.B., Safai, M., Eds.; International Society for Optics and Photonics, SPIE: Orlando, FL, USA, 2010; Volume 7661, pp. 49–58. [CrossRef]
33. Huang, D.; Wang, Y.; Wang, Y. A Robust Method for Near Infrared Face Recognition Based on Extended Local Binary Pattern. In *Proceedings of the Advances in Visual Computing*; Bebis, G., Boyle, R., Parvin, B., Koracin, D., Paragios, N., Tanveer, S.M., Ju, T., Liu, Z., Coquillart, S., Cruz-Neira, C., et al., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 437–446. [CrossRef]
34. Liu, L.; Zhao, L.; Long, Y.; Kuang, G.; Fieguth, P. Extended local binary patterns for texture classification. *Image Vis. Comput.* **2012**, *30*, 86–99. [CrossRef]
35. Zhao, G.; Huang, X.; Taini, M.; Li, S.Z.; Pietikäinen, M. Facial expression recognition from near-infrared videos. *Image Vis. Comput.* **2011**, *29*, 607–619. [CrossRef]
36. Hong, X.; Xu, Y.; Zhao, G. LBP-TOP: A Tensor Unfolding Revisit. In *Proceedings of the Computer Vision—ACCV 2016 Workshops*; Chen, C.S., Lu, J., Ma, K.K., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 513–527. [CrossRef]
37. Xie, Z. Infrared face recognition based on LBP co-occurrence matrix. In Proceedings of the 33rd Chinese Control Conference, Nanjing, China, 28–30 July 2014; pp. 4817–4820. [CrossRef]
38. Sujatha, B.; Kumar, V.; Harini, P. A new logical compact LBP co-occurrence matrix for texture analysis. *Int. J. Sci. Eng. Res.* **2012**, *3*, 1–5.

39. Lowe, D. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157. [CrossRef]
40. Yang, J.; Liao, S.; Li, S.Z. Automatic Partial Face Alignment in NIR Video Sequences. In Proceedings of the Advances in Biometrics; Tistarelli, M., Nixon, M.S., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 249–258. [CrossRef]
41. Zou, X.; Kittler, J.; Messer, K. Face Recognition Using Active Near-IR Illumination. In Proceedings of the British Machine Vision Conference, Oxford, UK, 5–8 September 2005. [CrossRef]
42. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. [CrossRef]
43. Friedrich, G.; Yeshurun, Y. Seeing People in the Dark: Face Recognition in Infrared Images. In *Proceedings of the Biologically Motivated Computer Vision Second International Workshop, BMCV 2002, Tübingen, Germany, 22–24 November 2002*; Bülthoff, H.H., Lee, S., Poggio, T.A., Wallraven, C., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2002; Volume 2525, pp. 348–359. [CrossRef]
44. Wu, S.Q.; Song, W.; Jiang, L.J.; Xie, S.L.; Pan, F.; Yau, W.Y.; Ranganath, S. Infrared Face Recognition by Using Blood Perfusion Data. In *Proceedings of the Audio- and Video-Based Biometric Person Authentication*; Kanade, T., Jain, A., Ratha, N.K., Eds.; Springer: Berlin/Heidelberg, Germany, 2005; pp. 320–328. [CrossRef]
45. Akhloufi, M.; Bendada, A. Thermal Faceprint: A New Thermal Face Signature Extraction for Infrared Face Recognition. In Proceedings of the 2008 Canadian Conference on Computer and Robot Vision, Windsor, ON, Canada, 28–30 May 2008; pp. 269–272. [CrossRef]
46. Huang, G.B.; Mattar, M.; Berg, T.; Learned-Miller, E. Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*; Erik Learned-Miller and Andras Ferencz and Frédéric Jurie: Marseille, France, 2008.
47. Zhang, Z.; Song, Y.; Qi, H. Age Progression/Regression by Conditional Adversarial Autoencoder. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 4352–4360. [CrossRef]
48. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 3730–3738. [CrossRef]
49. Wolf, L.; Hassner, T.; Maoz, I. Face recognition in unconstrained videos with matched background similarity. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 529–534. [CrossRef]
50. Nech, A.; Kemelmacher-Shlizerman, I. Level Playing Field for Million Scale Face Recognition. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3406–3415. [CrossRef]
51. Phillips, P.; Wechsler, H.; Huang, J.; Rauss, P.J. The FERET database and evaluation procedure for face-recognition algorithms. *Image Vis. Comput.* **1998**, *16*, 295–306. [CrossRef]
52. Marszalec, E.A.; Martinkauppi, J.B.; Soriano, M.N.; Pietikaeinen, M. Physics-based face database for color research. *J. Electron. Imaging* **2000**, *9*, 32–38. [CrossRef]
53. Li, S.Z.; Yi, D.; Lei, Z.; Liao, S. The CASIA NIR-VIS 2.0 Face Database. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013; pp. 348–353. [CrossRef]
54. Zhang, B.; Zhang, L.; Zhang, D.; Shen, L. Directional binary code with application to PolyU near-infrared face database. *Pattern Recognit. Lett.* **2010**, *31*, 2337–2344. [CrossRef]
55. Wang, S.; Liu, Z.; Lv, S.; Lv, Y.; Wu, G.; Peng, P.; Chen, F.; Wang, X. A Natural Visible and Infrared Facial Expression Database for Expression Recognition and Emotion Inference. *IEEE Trans. Multimed.* **2010**, *12*, 682–691. [CrossRef]
56. IRIS Thermal/Visible Face Database. Available online: http://vcipl-okstate.org/pbvs/bench/ (accessed on 14 September 2021).
57. Xu, Y.; Zhong, A.; Yang, J.; Zhang, D. Bimodal biometrics based on a representation and recognition approach. *Opt. Eng.* **2011**, *50*, 037202. [CrossRef]
58. Shoja Ghiass, R. Face Recognition Using Infrared Vision. Ph.D. Thesis, Université Laval, Quebec, Canada, 2018.
59. Kowalski, M.; Grudzień, A. High-resolution thermal face dataset for face and expression recognition. *Metrol. Meas. Syst.* **2018**, *25*, 403–415. [CrossRef]
60. Kopaczka, M.; Kolk, R.; Merhof, D. A fully annotated thermal face database and its application for thermal facial expression recognition. In Proceedings of the 2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Houston, TX, USA, 14–17 May 2018; pp. 1–6. [CrossRef]
61. Simón, M.O.; Corneanu, C.; Nasrollahi, K.; Nikisins, O.; Escalera, S.; Sun, Y.; Li, H.; Sun, Z.; Moeslund, T.B.; Greitans, M. Improved RGB-D-T based face recognition. *IET Biom.* **2016**, *5*, 297–303. [CrossRef]
62. Guo, K.; Wu, S.; Xu, Y. Face recognition using both visible light image and near-infrared image and a deep network. *CAAI Trans. Intell. Technol.* **2017**, *2*, 39–47. [CrossRef]
63. Flynn, P.J.; Bowyer, K.W.; Phillips, P.J. Assessment of Time Dependency in Face Recognition: An Initial Study. In *Proceedings of the Audio- and Video-Based Biometric Person Authentication*; Kittler, J., Nixon, M.S., Eds.; Springer: Berlin/Heidelberg, Germany, 2003; pp. 44–51. [CrossRef]
64. Akhloufi, M.A.; Bendada, A. A multistep approach for infrared face recognition in texture space. In *Proceedings of the Thermosense: Thermal Infrared Applications XXXV*; Stockton, G.R., Colbert, F.P., Eds.; International Society for Optics and Photonics, SPIE: Baltimore, MD, USA, 2013; Volume 8705, p. 87050C. [CrossRef]

65. Poster, D.; Thielke, M.; Nguyen, R.; Rajaraman, S.; Di, X.; Fondje, C.N.; Patel, V.M.; Short, N.J.; Riggan, B.S.; Nasrabadi, N.M.; et al. A Large-Scale, Time-Synchronized Visible and Thermal Face Dataset. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 1558–1567. [CrossRef]
66. Huang, D.; Sun, J.; Wang, Y. The Buaa-Visnir Face Database Instructions. *School Comput. Sci. Eng., Beihang Univ., Beijing, China, Tech. Rep. IRIP-TR-12-FR-001*. 2012. Available online: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=oqFMIuwAAAAJ&citation_for_view=oqFMIuwAAAAJ:qjMakFHDy7sC (accessed on 14 September 2021).
67. Bernhard, J.; Barr, J.; Bowyer, K.W.; Flynn, P. Near-IR to visible light face matching: Effectiveness of pre-processing options for commercial matchers. In Proceedings of the 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), Arlington, VA, USA, 8–11 September 2015; pp. 1–8. [CrossRef]
68. Hu, S.; Short, N.J.; Riggan, B.S.; Gordon, C.; Gurton, K.P.; Thielke, M.; Gurram, P.; Chan, A.L. A Polarimetric Thermal Database for Face Recognition Research. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 187–194. [CrossRef]
69. Szankin, M.; Kwasniewska, A.; Ruminski, J. Influence of Thermal Imagery Resolution on Accuracy of Deep Learning based Face Recognition. In Proceedings of the 2019 12th International Conference on Human System Interaction (HSI), Richmond, VA, USA, 25–27 June 2019; pp. 1–6. [CrossRef]
70. Espinosa-Duró, V.; Faundez-Zanuy, M.; Mekyska, J. A New Face Database Simultaneously Acquired in Visible, Near-Infrared and Thermal Spectrums. *Cogn. Comput.* **2013**, *5*, 119–135. [CrossRef]
71. Ariffin, S.M.Z.S.Z.; Jamil, N.; Rahman, P.N.M.A. Terravic Facial IR Database/IRIS Thermal/Visible Face Database/CBSR NIR Face Dataset. In *Proceedings of the 2016 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*; IEEE: Poznan, Poland, 2016; pp. 191–195. [CrossRef]
72. Dhamecha, T.I.; Nigam, A.; Singh, R.; Vatsa, M. Disguise detection and face recognition in visible and thermal spectrums. In Proceedings of the 2013 International Conference on Biometrics (ICB), Madrid, Spain, 4–7 June 2013; pp. 1–8. [CrossRef]
73. Wu, F.; You, W.; Smith, J.S.; Lu, W.; Zhang, B. Image-Image Translation to Enhance Near Infrared Face Recognition. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3442–3446. [CrossRef]
74. Panetta, K.; Wan, Q.; Agaian, S.; Rajeev, S.; Kamath, S.; Rajendran, R.; Rao, S.P.; Kaszowska, A.; Taylor, H.A.; Samani, A.; et al. A Comprehensive Database for Benchmarking Imaging Systems. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 509–520. [CrossRef]
75. Ashrafi, R.; Azarbayjani, M.; Tabkhi, H. Charlotte-ThermalFace: A Fully Annotated Thermal Infrared Face Dataset with Various Environmental Conditions and Distances. *Infrared Phys. Technol.* **2022**, *124*, 104209. [CrossRef]
76. Hanley, J.A. Receiver operating characteristic (ROC) methodology: The state of the art. *Crit. Rev. Diagn. Imaging* **1989**, *29*, 307–335. [PubMed]
77. Aggarwal, G.; Biswas, S.; Flynn, P.J.; Bowyer, K.W. Predicting performance of face recognition systems: An image characterization approach. In Proceedings of the CVPR 2011 WORKSHOPS, Colorado Springs, CO, USA, 20–25 June 2011; pp. 52–59. . [CrossRef]
78. Johnson, A.Y.; Sun, J.; Bobick, A.F. Predicting Large Population Data Cumulative Match Characteristic Performance from Small Population Data. In *Proceedings of the Audio- and Video-Based Biometric Person Authentication*; Kittler, J., Nixon, M.S., Eds.; Springer: Berlin/Heidelberg, Germany, 2003; pp. 821–829. [CrossRef]
79. Kumar, B.V.; Mahalanobis, A.; Song, S.; Sims, S.R.F.; Epperson, J.F. Minimum squared error synthetic discriminant functions. *Opt. Eng.* **1992**, *31*, 915–922. [CrossRef]
80. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A Discriminative Feature Learning Approach for Deep Face Recognition. In *Proceedings of the Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 499–515. [CrossRef]
81. Deng, Z.; Peng, X.; Li, Z.; Qiao, Y. Mutual Component Convolutional Neural Networks for Heterogeneous Face Recognition. *IEEE Trans. Image Process.* **2019**, *28*, 3102–3114. [CrossRef]
82. Deng, Z.; Peng, X.; Qiao, Y. Residual Compensation Networks for Heterogeneous Face Recognition. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 8239–8246. [CrossRef]
83. Xu, R.; Cho, M.; Lee, S. A NIR-to-VIS face recognition via part adaptive and relation attention module. *arXiv* **2021**, arXiv:2102.00689.
84. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
85. Lai, K.; Yanushkevich, S.N. Multi-Metric Evaluation of Thermal-to-Visual Face Recognition. In Proceedings of the 2019 Eighth International Conference on Emerging Security Technologies (EST), Colchester, UK, 22–24 July 2019; pp. 1–6. [CrossRef]
86. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
87. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. [CrossRef]
88. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.

89. Litvin, A.; Nasrollahi, K.; Escalera, S.; Ozcinar, C.; Moeslund, T.B.; Anbarjafari, G. A novel deep network architecture for reconstructing RGB facial images from thermal for face recognition. *Multimed. Tools Appl.* **2019**, *78*, 25259–25271. [CrossRef]
90. Quan, T.M.; Hildebrand, D.G.C.; Jeong, W.K. FusionNet: A Deep Fully Residual Convolutional Neural Network for Image Segmentation in Connectomics. *Front. Comput. Sci.* **2021**, *3*, 34. [CrossRef]
91. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier nonlinearities improve neural network acoustic models. *Proc. icml. Citeseer* **2013**, *30*, 3.
92. He, R.; Cao, J.; Song, L.; Sun, Z.; Tan, T. Cross-spectral Face Completion for NIR-VIS Heterogeneous Face Recognition. *arXiv* **2019**, arXiv:1902.03565.
93. Wu, X.; Huang, H.; Patel, V.M.; He, R.; Sun, Z. Disentangled Variational Representation for Heterogeneous Face Recognition. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 9005–9012. [CrossRef]
94. Wu, X.; He, R.; Sun, Z.; Tan, T. A Light CNN for Deep Face Representation with Noisy Labels. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 2884–2896. [CrossRef]
95. Guo, Y.; Zhang, L.; Hu, Y.; He, X.; Gao, J. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In *Proceedings of the Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 87–102. [CrossRef]
96. Guei, A.C.; Akhloufi, M.A. Deep generative adversarial networks for infrared image enhancement. In *Proceedings of the Thermosense: Thermal Infrared Applications XL*; Burleigh, D., de Vries, J., Eds.; International Society for Optics and Photonics, SPIE: Orlando, FL, USA, 2018; Volume 10661, p. 106610B. [CrossRef]
97. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv* **2015**, arXiv:1511.06434. [CrossRef]
98. Immidisetti, R.; Hu, S.; Patel, V.M. Simultaneous Face Hallucination and Translation for Thermal to Visible Face Verification using Axial-GAN. In Proceedings of the 2021 IEEE International Joint Conference on Biometrics (IJCB), Shenzhen, China, 4–7 August 2021; pp. 1–8. [CrossRef]
99. Kim, J.; Ra, M.; Kim, W.Y. A DCNN-Based Fast NIR Face Recognition System Robust to Reflected Light From Eyeglasses. *IEEE Access* **2020**, *8*, 80948–80963. [CrossRef]
100. Luo, Y.; Pi, D.; Pan, Y.; Xie, L.; Yu, W.; Liu, Y. ClawGAN: Claw connection-based generative adversarial networks for facial image translation in thermal to RGB visible light. *Expert Syst. Appl.* **2022**, *191*, 116269. [CrossRef]
101. Wu, Z.; Peng, M.; Chen, T. Thermal face recognition using convolutional neural network. In Proceedings of the 2016 International Conference on Optoelectronics and Image Processing (ICOIP), Warsaw, Poland, 10–12 June 2016; pp. 6–9. [CrossRef]
102. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893. [CrossRef]
103. de Freitas Pereira, T.; Anjos, A.; Marcel, S. Heterogeneous Face Recognition Using Domain Specific Units. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 1803–1816. [CrossRef]
104. Peng, M.; Wang, C.; Chen, T.; Liu, G. NIRFaceNet: A Convolutional Neural Network for Near-Infrared Face Identification. *Information* **2016**, *7*, 61. [CrossRef]
105. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper With Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015. [CrossRef]
106. Hu, W.; Hu, H. Disentangled Spectrum Variations Networks for NIR–VIS Face Recognition. *IEEE Trans. Multimed.* **2020**, *22*, 1234–1248. [CrossRef]
107. Kim, J.; Jo, H.; Ra, M.; Kim, W.Y. Fine-tuning Approach to NIR Face Recognition. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2337–2341. [CrossRef]
108. Shavandi, M.; Afrakoti, I.E.P. Face Recognition in Thermal Images based on Sparse Classifier. *Int. J. Eng.* **2019**, *32*, 78–84. [CrossRef]
109. Mahouachi, D.; Akhloufi, M.A. Adaptive deep convolutional neural network for thermal face recognition. In *Proceedings of the Thermosense: Thermal Infrared Applications XLIII*; Zalameda, J.N., Mendioroz, A., Eds.; International Society for Optics and Photonics, SPIE: Online Only, 2021; Volume 11743, p. 1174304. [CrossRef]
110. Mahouachi, D.E.; Akhloufi, M.A. Deep adaptive convolutional neural network for near infrared and thermal face recognition. In *Proceedings of the Infrared Technology and Applications XLVIII*; Andresen, B.F., Fulop, G.F., Zheng, L., Eds.; International Society for Optics and Photonics, SPIE: Orlando, FL, USA, 2022; Volume 12107, p. 121071R. [CrossRef]
111. Jo, H.; Kim, W.Y. NIR Reflection Augmentation for DeepLearning-Based NIR Face Recognition. *Symmetry* **2019**, *11*, 1234. [CrossRef]
112. Gavini, Y.; Mehtre, B.M.; Agarwal, A. Thermal to Visual Face Recognition using Transfer Learning. In Proceedings of the 2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA), Hyderabad, India, 22–24 January 2019; pp. 1–8. [CrossRef]
113. Wang, P.; Bai, X. Regional parallel structure based CNN for thermal infrared face identification. *Integr.-Comput.-Aided Eng.* **2018**, *25*, 247–260. [CrossRef]

114. He, R.; Wu, X.; Sun, Z.; Tan, T. Wasserstein CNN: Learning Invariant Features for NIR-VIS Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1761–1773. [CrossRef]
115. Lezama, J.; Qiu, Q.; Sapiro, G. Not Afraid of the Dark: NIR-VIS Face Recognition via Cross-Spectral Hallucination and Low-Rank Embedding. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6807–6816. [CrossRef]
116. Cho, M.; Kim, T.; Kim, I.J.; Lee, K.; Lee, S. Relational Deep Feature Learning for Heterogeneous Face Recognition. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 376–388. [CrossRef]
117. Kumar, S.; Singh, S.K. Occluded Thermal Face Recognition Using Bag of CNN (*Bo*CNN). *IEEE Signal Process. Lett.* **2020**, *27*, 975–979. [CrossRef]
118. Sarfraz, M.S.; Stiefelhagen, R. Deep Perceptual Mapping for Thermal to Visible Face Recognition. *arXiv* **2015**, arXiv:1507.02879. [CrossRef]
119. He, R.; Wu, X.; Sun, Z.; Tan, T. Learning Invariant Deep Representation for NIR-VIS Face Recognition. *Proc. AAAI Conf. Artif. Intell.* **2017**, *31*. [CrossRef]
120. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
121. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 Ocotber 2017; pp. 2242–2251. [CrossRef]
122. Wang, Y.; Li, Y.; Wang, S. Parallel-Structure-based Transfer Learning for Deep NIR-to-VIS Face Recognition. In *Proceedings of the Image and Graphics*; Zhao, Y., Barnes, N., Chen, B., Westermann, R., Kong, X., Lin, C., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 146–156. [CrossRef]
123. Liu, X.; Song, L.; Wu, X.; Tan, T. Transferring deep representation for NIR-VIS heterogeneous face recognition. In Proceedings of the 2016 International Conference on Biometrics (ICB), Halmstad, Sweden, 13–16 June 2016; pp. 1–8. [CrossRef]
124. Zhao, P.; Zhang, F.; Wei, J.; Zhou, Y.; Wei, X. SADG: Self-Aligned Dual NIR-VIS Generation for Heterogeneous Face Recognition. *Appl. Sci.* **2021**, *11*, 987. [CrossRef]
125. Hu, W.; Hu, H. Dual Adversarial Disentanglement and Deep Representation Decorrelation for NIR-VIS Face Recognition. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 70–85. [CrossRef]
126. Sun, R.; Shan, X.; Zhang, H.; Gao, J. Data gap decomposed by auxiliary modality for NIR-VIS heterogeneous face recognition. *IET Image Process.* **2022**, *16*, 261–272. [CrossRef]
127. Cheema, U.; Ahmad, M.; Han, D.; Moon, S. Heterogeneous Visible-Thermal and Visible-Infrared Face Recognition Using Cross-Modality Discriminator Network and Unit-Class Loss. *Comput. Intell. Neurosci.* **2022**, *2022*, 4623368. [CrossRef] [PubMed]
128. Menon, S.; J., S.; S.K., A.; Nair, A.P.; S., S. Driver Face Recognition and Sober Drunk Classification using Thermal Images. In Proceedings of the 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 4–6 April 2019; pp. 400–404. [CrossRef]
129. M., S.K.K.; Rajendran, R.; Wan, Q.; Panetta, K.; Agaian, S.S. TERNet: A deep learning approach for thermal face emotion recognition. In *Proceedings of the Mobile Multimedia/Image Processing, Security, and Applications 2019*; Agaian, S.S., Asari, V.K., DelMarco, S.P., Eds.; International Society for Optics and Photonics, SPIE: Baltimore, FL, USA, 2019; Volume 10993, p. 1099309. [CrossRef]
130. Mohamed, S.; Ghoneim, A.; Youssif, A. Visible/Infrared face spoofing detection using texture descriptors. *MATEC Web Conf.* **2019**, *292*, 04006. [CrossRef]
131. Du, H.; Shi, H.; Liu, Y.; Zeng, D.; Mei, T. Towards NIR-VIS Masked Face Recognition. *IEEE Signal Process. Lett.* **2021**, *28*, 768–772. [CrossRef]
132. Wang, X.; Tang, X. Face Photo-Sketch Synthesis and Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 1955–1967. [CrossRef] [PubMed]
133. Chingovska, I.; Erdogmus, N.; Anjos, A.; Marcel, S., Face Recognition Systems Under Spoofing Attacks. In *Face Recognition Across the Imaging Spectrum*; Springer International Publishing: Cham, Switzerland, 2016; pp. 165–194. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.