applied sciences

MDPI

Article

SADG: Self-Aligned Dual NIR-VIS Generation for Heterogeneous Face Recognition

Pengcheng Zhao [1,2], Fuping Zhang [2,3], Jianming Wei [2,*], Yingbo Zhou [1,2] and Xiao Wei [1]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China; zhaopengcheng2018@sari.ac.cn (P.Z.); zhouyingbo2018@sari.ac.cn (Y.Z.); xwei@shu.edu.cn (X.W.)
[2] Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China; zhangfp@sari.ac.cn
[3] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: wjm@sari.ac.cn

Abstract: Heterogeneous face recognition (HFR) has aroused significant interest in recent years, with some challenging tasks such as misalignment problems and limited HFR data. Misalignment occurs among different modalities' images mainly because of misaligned semantics. Although recent methods have attempted to settle the low-shot problem, they suffer from the misalignment problem between paired near infrared (NIR) and visible (VIS) images. Misalignment can bring performance degradation to most image-to-image translation networks. In this work, we propose a self-aligned dual generation (SADG) architecture for generating semantics-aligned pairwise NIR-VIS images with the same identity, but without the additional guidance of external information learning. Specifically, we propose a self-aligned generator to align the data distributions between two modalities. Then, we present a multiscale patch discriminator to get high quality images. Furthermore, we raise the mean landmark distance (MLD) to test the alignment performance between NIR and VIS images with the same identity. Extensive experiments and an ablation study of SADG on three public datasets show significant alignment performance and recognition results. Specifically, the Rank1 accuracy achieved was close to 99.9% for the CASIA NIR-VIS 2.0, Oulu-CASIA NIR-VIS and BUAA VIS-NIR datasets, respectively.

Keywords: heterogeneous face recognition; NIR-VIS data generation; semantic alignment

Citation: 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. https://doi.org/10.3390/app11030987

Received: 21 December 2020
Accepted: 20 January 2021
Published: 22 January 2021

1. Introduction

Face recognition as a biometric identification method has attracted a large number of researchers' focus through the past several decades for its widespread applications and contactless acquisition [1]. Nevertheless, face recognition primarily concentrates on the visible spectrum, which brings about the more difficult problem of recognizing a subject under low-light or no-light conditions. Consequently, heterogeneous face recognition (HFR) research has emerged in recent years. Heterogeneous face images acquired under different spectra are different in their modalities, such as sketch images [2], near infrared (NIR) images [3] and polarimetric thermal images [4]. HFR with NIR images is an important computer vision task [5].

The literature over the years on HFR can mainly be divided into three branches: feature representation-based methods, common subspace projection-based methods and synthesis-based methods. Feature representation-based methods reduce the domain gap between heterogeneous images through handcrafted features such as linear discriminant analysis (LDA) and principal component analysis (PCA) [6]. Common subspace projection-based methods project different modality images into a common subspace [5]. However, there is a lack of a large-scale near infrared-visible (NIR-VIS) face database relative to visible face datasets, which boosts the research of synthesis-based methods [7]. They

usually transform NIR images to identity-preserved VIS ones and then evaluate models' recognition performance on the generated VIS images. Fu et al. [8] creatively proposed a new insight for HFR that utilized Gaussian noise to generate abundant paired NIR-VIS images through a dual variational generation (DVG) model. It covered the few-shot problem of HFR and improved the recognition performance of a limited dataset.

However, it remains a challenging problem for the synthesis-based methods that the paired NIR and VIS images from the same identity are semantically misaligned from each other. The image translation network will produce unsatisfying results, such as the variance of poses and expressions between the input and the output when training with unaligned data, which can degrade the recognition performance [9]. The misalignment problem is mainly caused by acquiring NIR and VIS images under different scenarios at different times, which brings variations in poses, expressions and so on, as shown in Figure 1a. In [3], the authors tried to resolve the misalignment problem with external face shape information for guidance, but it was still a two-stage image-to-image translation network and could not produce new face images with different identities beyond the original NIR-VIS datasets. Note that the model proposed by [8] cannot produce aligned, paired NIR-VIS images because of the misalignment problem in the original datasets. The raw data vary from each other due to complicated face attributes such as the pose, expression, hair and wearing glasses or not [10]. The autoencoder learns the misaligned distribution with the paired images and decodes them with a reconstruction constraint, hence leading to a semantically image-to-image network. Therefore, it is hard to produce paired NIR-VIS images with the same semantic information.

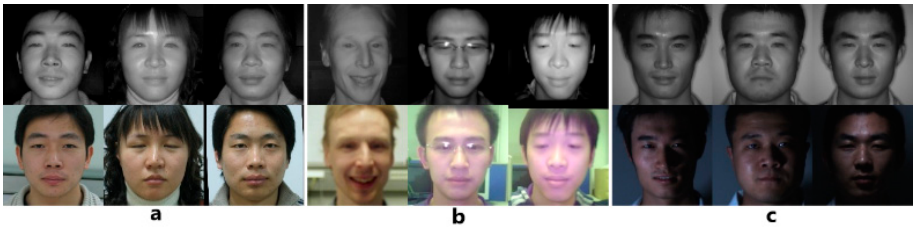

Figure 1. Paired near infrared-visible (NIR-VIS) images in three datasets (**a**–**c**) for the CASIA NIR-VIS 2.0, Oulu-CASIA NIR-VIS and BUAA VIS-NIR datasets, respectively.

To cover the misalignment problem, we propose a self-aligned generation architecture to align the data distributions between two modalities semantically. Specifically, we use two encoders and two decoders for generating paired images with different domains. A training method is promoted with the same latent code and a self-aligned block to train our network. The same latent code can affect the alignment performance virtually while the self-aligned block acts the part for redressing the unaligned attributes subsidiarily. These strategies guarantee the alignment of images between two domains. After the training stage, we can not only utilize our model to generate abundant NIR-VIS images with the same noise from a standard Gaussian distribution, but also redress the misaligned raw datasets with our well-trained model by reconstructing them. Furthermore, we present a multiscale patch discriminator for the high quality of the generated aligned NIR-VIS images. In the evaluation part, a new metric method, namely the mean landmark distance (MLD), is raised to test the alignment effect of the generated NIR and VIS images with the same identity. We train our model with the CASIA NIR-VIS 2.0 dataset [3] and our numerous generated NIR-VIS images, then evaluate the model performance on the Oulu-CASIA NIR-VIS [11] and BUAA VIS-NIR [12] datasets. The work of the self-aligned dual generation (SADG) method will be available at https://github.com/Renrenren6666/SADG.

In summary, our contributions are as follows:

1. We analyze the mechanism of the semantics misalignment problem between two modality images;
2. A self-aligned dual generation architecture (SADG) is proposed to align NIR and VIS images, including a self-aligned block and a multiscale patch discriminator;

3. The mean landmark distance (MLD) is raised to measure the alignment performance. Extensive experiments and an ablation study conducted on three popular datasets show the state-of-the-art alignment and recognition performance of our method.

2. Related Works

For HFR, the main focus is to reduce the domain gap. Generally, the literature over the years can be categorized into three groups: feature representation-based methods, common subspace projection-based methods and synthesis-based methods.

Feature representation-based methods mainly explore illumination invariance through hand-crafted features to reduce the modality gap. The earliest work on NIR-to-VIS face recognition was proposed by Yi et al. [6], who utilized PCA, LDA and canonical correlation analysis (CCA) in three steps to gain a better performance. Sarfraz and Stiefelhagen [13] explained the NIR-VIS problem as a task with high illumination variation. They solved this problem by designing an effective light invariant descriptor: a logarithmic gradient histogram (LGH). LGH was superior to the local binary pattern (LBP) and scale-invariant feature transform (SIFT) descriptors used in [14,15], as it was a pure function-based approach without training data. These methods could extract features from images of different modalities from the standpoint that they removed the domain information. However, they could not reach satisfactory recognition performance in most cases, which could only achieve a Rank1 accuracy of 70–80%.

Common subspace projection-based methods try to map heterogeneous face images into a latent common subspace, where the images in different domains can be matched straightaway. Lin and Tang [16] primarily proposed the common discriminant feature extraction (CDFE) approach, which could extract the recognition information and location information simultaneously. Kan et al. [17] raised the multi-view discriminant analysis (MVDA) method to study both the interview and intraview pertinence of heterogeneous face images. Huo et al. [18] presented a form of margin-based cross-modality metric learning to reduce the gaps of different modalities. For better extracting the discriminative information, a regularized discriminative spectral regression method was developed to find a common spectral space in [19], where it used the locality information in kernel space for discrimination. An extreme learning machine (ELM) combined with a multitask cluster were used by [20] for cross-model feature learning. Similar to the feature representation-based methods, these methods cannot achieve high performance for HFR, which can only achieve a Rank1 accuracy of about 90%.

Synthesis-based methods, with the development of deep learning [21] and generation networks such as generative adversarial networks (GANs) [22] and variational autoencoders (VAEs) [23], have aroused great interest among researchers. They usually translate NIR face images to identity-preserved VIS ones, and then evaluate the generated VIS images to match with the NIR ones. Riggan et al. [24] trained a regression network to estimate the projection between the features of the visible and thermal modalities, and then reconstructed the visible face image from the estimated features. Zhang et al. [25,26] leveraged GANs to synthesize visible images from the polarimetric thermal images. Recently, a novel network for generating large-scale paired NIR-VIS images by noise, which brought novel insight for HFR, was adopted [8]. Duan et al. [27] proposed a pose aligned cross-spectral hallucination (PACH) algorithm, which dealt with the facial shape and the spectrum information in two individual stages. Yu et al. [9] proposed a pose-preserving cross-spectral face hallucination (PCFH) model to synthesize a VIS image with the same identity as the given NIR image while preserving poses and expressions. However, they remain under the image-to-image translation framework.

As for the metrics in HFR, the recent literature can be mainly divided into two types: common subspace-based measurement function and bilinearity-based similarity measurement function. Common subspace-based measurement function compares two domains' features in the same space, which are obtained by a common subspace projection. Wu et al. [28] proposed a measurement function to evaluate the distance between two

domains' features, mapping the code features and text features into the same semantic space. Siena, Boddeti and Kumar [29] proposed a method for HFR named maximizing margin coupled mappings (MMCM), which narrowed the gap between samples of the same subject and increased the distance between samples of different ones in a pair. The similarity measurement function evaluates the domain gap by the similarity of features from different modalities. Zhen et al. [30] put forward a probability learning framework to distinguish the similar images from different domains. However, these methods focus on the domain gap in the feature level. They cannot measure the alignment effect of paired heterogeneous images with the same image level identity.

3. Method

As mentioned above, a novel dual variational generation framework to reduce the domain gap was proposed in [8], generating massive paired NIR-VIS images with the same identity for a pair by noise. However, it suffered from the misalignment problem. We utilize the backbone of DVG as a baseline model and improve it for alignment purposes. In this part, we will briefly introduce the baseline model first, and then analyze the mechanism of DVG as it pertains to the misalignment. Finally, we will dwell on the details of our method and discuss our improved architecture.

3.1. Baseline

As proposed in [8], the baseline network mainly consisted of a dual variational autoencoder. In particular, the autoencoder included two encoder networks E_N and E_V, as well as a decoder network D. Given the pairwise NIR-VIS images $\{x_N, x_V\}$, the encoder E_N plays the role of mapping NIR images x_N to the latent space z_N with a reparameterization trick: $z_N = \mu_N + \sigma_N \odot \tau$. μ_N and σ_N are defined as the mean and standard deviation of the feature maps from the NIR images in a mini-batch, respectively, τ is sampled from a multivariate standard Gaussian and $\odot$ denotes the Hadamard product. E_V conducts the same operation correspondingly and finally gets the latent code of the VIS images z_V. Note that this encoding process can be found in our architecture, as shown in Figure 2. Finally, the latent code z_N and z_V are concatenated to get a joint distribution z_I, which is utilized to reconstruct the input images by one decoder network in the baseline network.

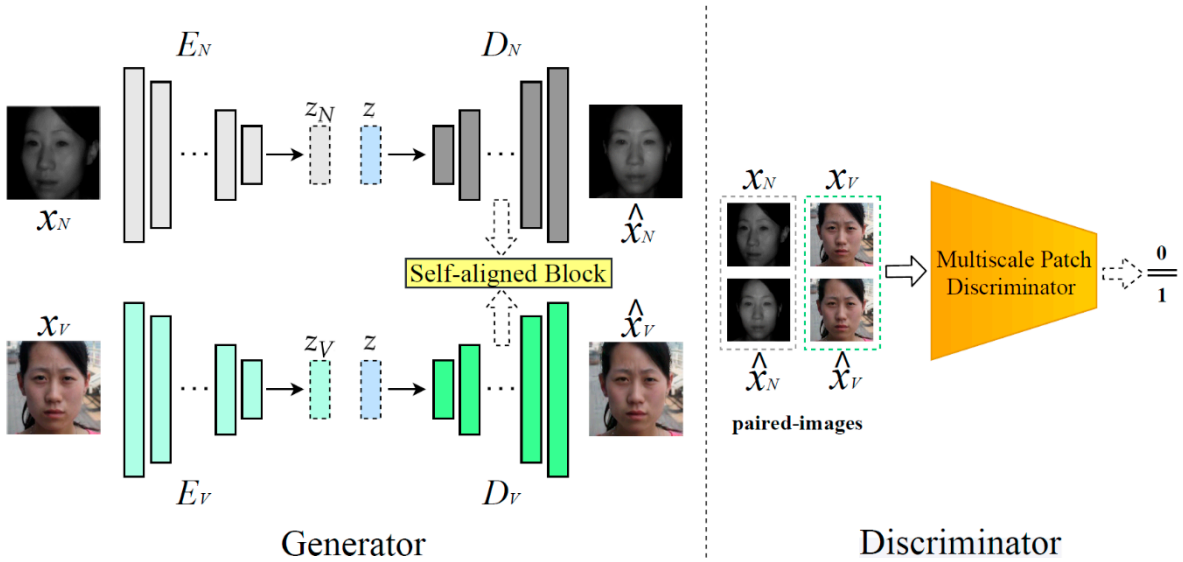

Figure 2. The training process of our proposed self-aligned generation architecture. Given the paired data from the training dataset, we first reverse them by E_N, D_N, E_V and D_V, respectively (the left part). Then, we use a multiscale patch discriminator (the right part) to do the adversarial learning with the generator. The solid-lined rectangle stands for the residual block, and the dotted-lined rectangle stands for the latent code z.

As with the VAE in the original work [23], DVG constrains the learning process of the posterior distributions $p_{\varnothing N}(z_N|x_N)$ and $p_{\varnothing V}(z_V|x_V)$ by a *KL* divergence:

$$L_{kl} = D_{KL}(p_{\varnothing N}(z_N|x_N)||p(z_N)) + D_{KL}(p_{\varnothing V}(z_V|x_V)||p(z_V)) \quad (1)$$

where the prior distributions $p(z_N)$ and $p(z_V)$ are both the multivariate standard Gaussian distributions. Then, a reconstruction loss is used to force the decoder network to reconstruct the input images $\{x_N, x_V\}$ from the learned distribution, which is modified in our method for alignment. We will discuss this in Section 3.2.

Except for L_{kl}, a simplified Wasserstein distance [8] between the two distributions $p(z_N^{(i)}) = N(\mu_N^{(i)}, \sigma_N^{(i)})$ and $p(z_V^{(i)}) = N(\mu_V^{(i)}, \sigma_V^{(i)})$ is also utilized for posterior distribution learning in latent space, where i stands for the identity:

$$L_{dist} = \sum_i \frac{1}{2}[\|\mu_N^{(i)} - \mu_V^{(i)}\|_2^2 + \|\sigma_N^{(i)} - \sigma_V^{(i)}\|_2^2] \quad (2)$$

For the image space, DVG uses a pre-trained LightCNNv2 [31] model as the Identity-feature extractor F_{id} for calculating identification loss. Due to DVG producing a pair of NIR-VIS images for each allotment of time, we can retain the identity information by constraining the feature distance between the reconstructed paired images. The same operation is conducted for the reconstructed image with the original image, and the loss functions are as follows:

$$L_{id\text{-}pair} = \|F_{id}(\hat{X}_N) - F_{id}(\hat{X}_V)\|_2^2 \quad (3)$$

$$L_{id\text{-}recpair} = \|F_{id}(\hat{X}_N) - F_{id}(x_N)\|_2^2 + \|F_{id}(\hat{X}_V) - F_{id}(x_V)\|_2^2 \quad (4)$$

where $F(\cdot)$ denotes the uniformed output of the last fully connected layer in F_{id}. For the purpose of increasing the diversity of the generated images, a diversity loss [32], defined as L_{div}, is adopted as the original model [8].

After training the network, DVG can generate paired NIR-VIS images with the same identity by the same noise sampled from Gaussian space, which can boost the limited dataset.

3.2. Architecture

As described above, DVG learns the data distribution from the training dataset, which suffers from the semantic misalignment between two different modalities, as shown in Figure 1. Therefore, the misalignment occurs first in the latent code z_N and z_V during the encoding process. Note that L_{dist} in Equation (2) decreases the Wasserstein distance of the two distributions representing different modalities, which can only maintain the identity information to achieve the same semantic direction with the same identity. The misaligned latent code z_N and z_V are then concatenated as z_I as the input of the decoder. As such, it is challengeable for DVG to generate semantically aligned NIR-VIS images with the same noise.

Recently, precisely semantic face editing with manipulation in the latent space code z was explored, in which we can edit one specific face attribute by a linear operation to change the semantic direction $z' = z + z_0$ [33]. z_0 is the direction of a certain face attribute, and then the generator can create attribute-changed images from z'. It indicates the semantic information in a code can be redressed with some specific operation. However, we cannot make the paired NIR-VIS images aligned by the same operation in our scene because the original paired images are misaligned in diversity attributes, such as poses with different angles, different expressions or wearing or not wearing glasses. That is to say, we cannot align the semantic distribution of two modalities' images in the learned latent distribution by E_N and E_V due to different identities having different misaligned face attributes. We cannot align all the code from different paired heterogeneous face images with one simple operation.

Therefore, we modify the baseline and propose a new architecture for generating aligned pairwise NIR-VIS images with the same identity. To be specific, we improve the generation process and adversarial learning part in our architecture, which includes a self-aligned generator and a multiscale patch discriminator, as shown in Figure 2.

3.2.1. Self-Aligned Generator

Inspired by [7,33,34], we propose a self-aligned generator for producing ID-consistent aligned NIR-VIS images, including a self-aligned block. As shown in the left part of Figure 2, we utilized two decoders, D_N and D_V, with the same structure in our generator, which was different from the baseline. Note that the architecture of our decoders was the same as that in the baseline.

As for the training stage, given pairwise NIR and VIS images $\{x_N, x_V\}$, we also encoded them to latent code z_N and z_V as a baseline. L_{kl} and L_{dist} in Equations (1) and (2) were used for learning the data distribution. Considering the semantic deviation between z_N and z_V, we sent the same code to D_N and D_V from one domain at the same time for reconstructing the input paired images x_N and x_V, which could align the semantic information in the code level. We further aligned one domain's semantic information to another by a self-aligned block in the feature level. Note that z in Figure 2 is one code from z_N or z_V. We used z_V for exhibition in the experimental part.

Figure 3 shows our self-aligned block, mainly made up of a self-attention module [35]. First of all, we fetched out the feature maps from the same layer in decoders D_α and D_β. Secondly, we fed features from one domain into the self-attention module. Then, the feature maps were transformed into three parts—a, b and c—with two 1×1 convolutional layers, which halved the channels of features a and b. Next, parts a and b performed elementwise multiplication, followed by a 1×1 convolutional layer and a softmax operation to get the attention maps m. Finally, $f_1^{(l)}$ was exported from this block through the elementwise multiplication of the feature maps c and attention maps m. The feature maps of other domains in the lth layer were denoted as $f_2^{(l)}$. We used the Euclidean distance to reduce the semantic gap of the feature maps between two domains:

$$L_a = \sum_{l=n}^{Y} \| f_1^{(l)} - f_2^{(l)} \|_2^2 \tag{5}$$

where n and Y stand for the number of residual blocks in the two decoders. Note that we used the self-aligned block in the final convolution layer in last three residual blocks. With the help of the self-aligned block, we could extract the most important semantic information in one domain and align another one's semantic information to it. We used it in multiscale feature maps in several layers of the two decoders for better alignment performance. Regarding the reconstructed images, we employed a reconstruction loss [36] with a low weight for domain α and a normal weight for domain β. Specifically, we needed the decoders $logp_\theta(x_\alpha|z_\alpha)$ and $logq_\theta(x_\beta|z_\beta)$ to reconstruct the input images $\{x_N, x_V\}$ to a different extent:

$$L_{rec} = -E_{p_{\varnothing\alpha(z_\alpha|x_\alpha)}} logp_\theta(x_\alpha|z_\alpha) - \rho E_{q_{\varnothing\beta(z_\beta|x_\beta)}} logq_\theta\left(x_\beta|z_\beta\right) \tag{6}$$

where ρ is a very low weight for the images x_α from domain α and x_β is the image to be aligned, whose domain is β. Note that the self-attention block and the code z should have been adjusted according to the aligned modality. For example, if we wanted to align the semantic information in the NIR modality to the VIS one, we should send the same VIS code z_V to D_N and D_V first. Then, we extracted the significant semantic information in the VIS modality by our self-attention block and use L_a to constrain the semantic gap between rgw VIS and NIR modalities. Finally, we used the reconstruction loss L_{rec} to force the NIR images to maintain their domain information at the image level and force the VIS images to be rebuilt faithfully. In this way, we could generate aligned NIR-VIS images, where the

created NIR images were aligned to the VIS images semantically. Note that both of the codes from two domains were equal in the training stage.

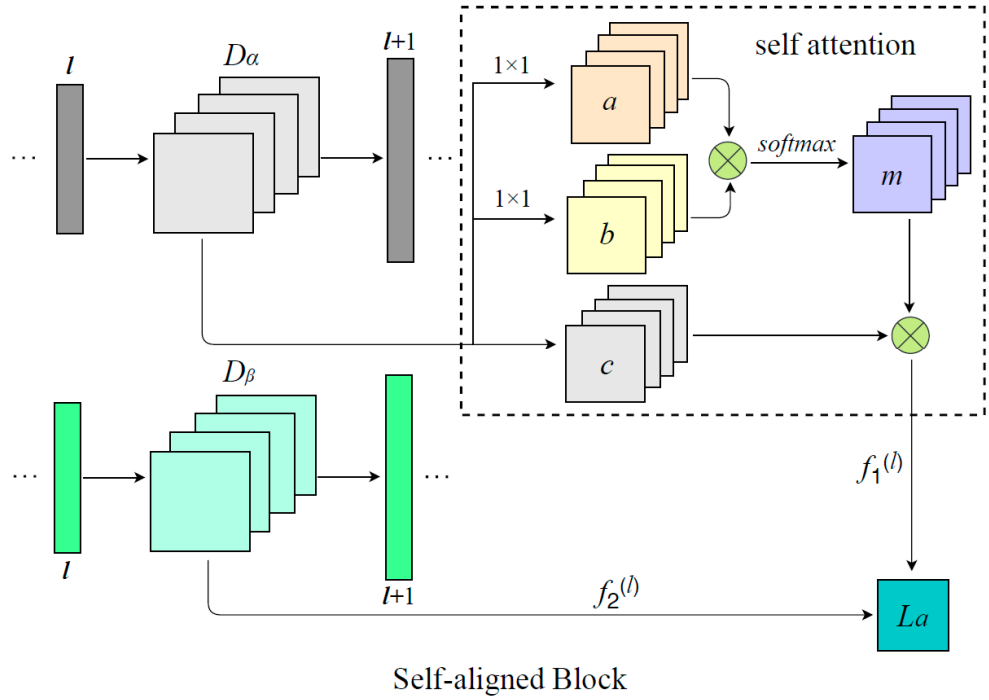

Figure 3. The architecture of the self-aligned block. The rectangle stands for residual block in D_α and D_β. The square stands for the feature maps. α and β stand for the two different domains. Images from the domain β are the images to be aligned.

We adopted $L_{id\text{-}pair}$ and $L_{id\text{-}recpair}$ in Equations (3) and (4) to restrict the identity consistency, as [8] did. To be specific, $L_{id\text{-}pair}$ restricts the identity consistency between the generated paired images with different modalities, and $L_{id\text{-}recpair}$ makes the reconstructed data have the same identity with the raw data. After the reconstruction procedure, we sent the original data $\{x_N, x_V\}$ and the reconstructed data $\{\hat{X}_N, \hat{X}_V\}$ to our multiscale patch discriminator for adversarial learning, which will be demonstrated in the next section minutely.

3.2.2. Multiscale Patch Discriminator

We propose a multiscale patch discriminator to promote the quality of generated pairwise NIR-VIS images. Similar to [37], a patch discriminator was used as our backbone to distinguish the authenticity of the input images, which consisted of several convolutional blocks followed by batch normalization and Leaky ReLU operations. We inserted these convolutional blocks before the final fully connected layer to get the multiscale feature maps, which can help to improve the capability of the discriminator.

As shown in Figure 4, in the training stage, we first matched the raw data and reconstructed data with the same corresponding modality and sent them into the feature extractor part of the network to get the feature maps. Different from [37], we then used four different convolutional blocks to get the final feature maps with different sizes, followed by a sigmoid function to get the confidence scores normalized between 0 and 1. The adversarial loss is defined as follows:

$$L_{aN}(X_N, \hat{X}_N, D_p) = E_{\hat{X}_N \sim \hat{X}_N}\left[\sum_{p=1}^{4} log(D_p(\hat{X}_N))\right] + E_{x_N \sim X_N}[\sum_{p=1}^{4}(1 - log(D_p(x_N)))] \quad (7)$$

$$L_{aV}(X_V, \hat{X}_V, D_p) = E_{\hat{X}_V \sim \hat{X}_V}\left[\sum_{p=1}^{4} log(D_p(\hat{X}_V))\right] + E_{x_V \sim X_V}[\sum_{p=1}^{4}(1 - log(D_p(x_V)))] \quad (8)$$

where p is the number of the multiscale patch discriminator and $\hat{X}$, X, $\hat{x}$, x, N and V indicate the reconstructed images set, the raw images set, one reconstructed image, one raw image, the NIR domain and the VIS domain, respectively. We designed the spatial sizes of the final feature maps as $1/2$, $1/4$, $1/8$ and $1/16$ of those of the input images, respectively.

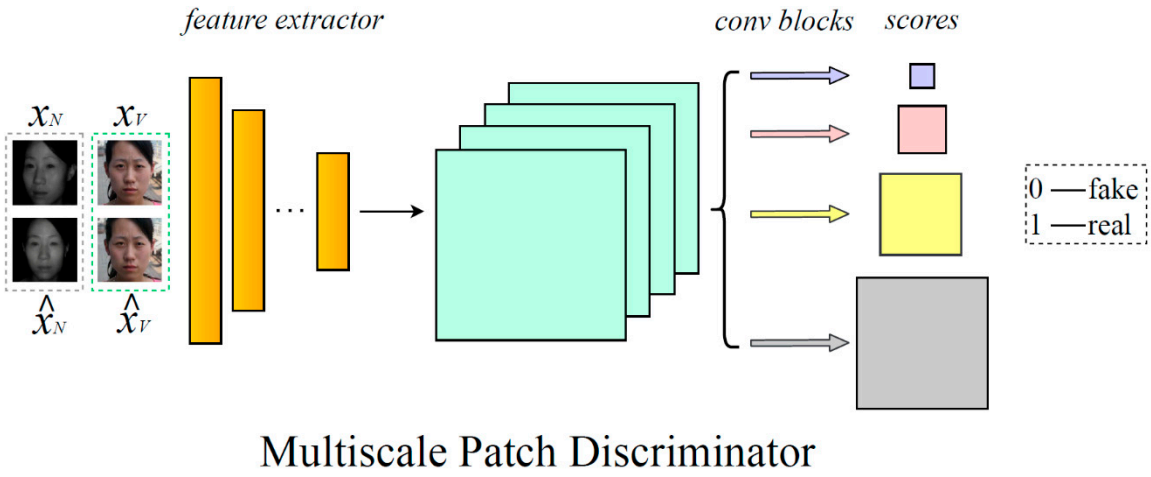

Figure 4. The architecture of the multiscale patch discriminator. The rectangle stands for the convolutional block in the feature extractor, and the arrowheads with different colors stand for the different convolutional blocks for getting final feature maps with different sizes.

Hence, the total loss for training our generation model can be formulated as

$$L = L_{kl} + \alpha_1 L_{dist} + \alpha_{2L_{id\text{-}pair}} + \alpha_3 L_{id\text{-}recpair} + \alpha_4 L_{div} + \alpha_5 L_{rec} + L_a + L_{aN} + L_{aV} \quad (9)$$

where α_1, α_2, α_3, α_4 and α_5 are the trade-off parameters. After training our model, we could use two decoders—D_N and D_V—to produce numerous pairwise aligned NIR-VIS images, with the same noise sampled in Gaussian space for a pair, while keeping the identity consistency between them. The details of the generation process are presented in Figure 5.

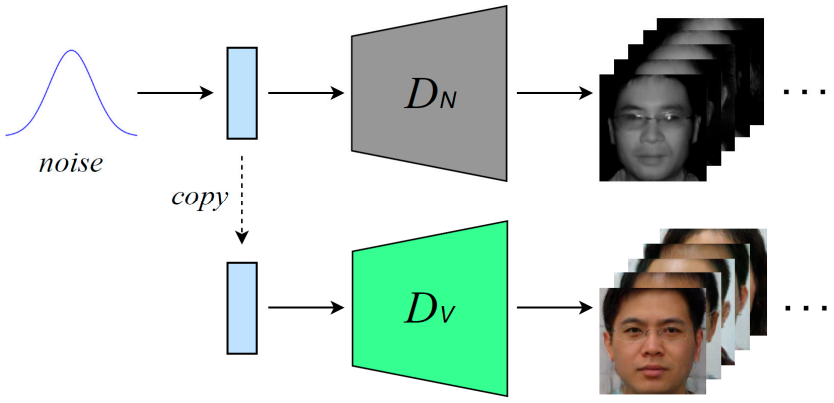

Figure 5. The process of generating well-aligned near infrared-visible (NIR-VIS) images with the same identity.

3.3. NIR-VIS Face Recognition

We followed the baseline [8] to adopt the LightCNN method F for NIR-VIS face recognition. We trained F with both the limited, original labeled data and the plentiful,

unlabeled aligned NIR-VIS images, which were generated by our well-trained generator. Quantitative analysis will be conducted for comparison in Section 4.

3.4. Mean Landmark Distance

We proposed the mean landmark distance (MLD) to quantitatively evaluate the alignment effect of two NIR-VIS images with the same identity. To be specific, we employed two facial landmark localization models named Landmark-5 [38] and Landmark-68 [39] for facial keypoint localization, both of which are widely used in face detection and face recognition tasks. Landmark-5 can detect 5 keypoints' coordinates, which locate the mouth, eyes and nose, while Landmark-68 can detect 68 keypoints' coordinates, which locate the mouth, eyes, nose and face contour. As for each paired set of NIR-VIS images, we computed the mean coordinate deviation for every keypoint in the face between two domains first. Then, we did the same operation with 100,000 pairwise generated fake images and took the average value as our MLD. MLD5 was computed by Landmark-5, which could stand for the alignment of the facial organs, and MLD68 was computed by Landmark-68, which could stand for not only the alignment of the facial organs, but also the alignment of the face contour:

$$\mathrm{MLD} = \frac{1}{nK}\sum_{k=1}^{K}\left(\left|l_{x_N}^{(k)} - l_{x_V}^{(k)}\right| + \left|l_{y_N}^{(k)} - l_{y_V}^{(k)}\right|\right) \tag{10}$$

where n stands for the number of generated paired images, $l_{x_N}^{(k)}$ is denoted as the kth keypoint's x-coordinate of the NIR image and $l_{x_V}^{(k)}$ is denoted as the kth keypoint's x-coordinate of the VIS image. It is the same with the y-coordinate in $l_{y_N}^{(k)}$ and $l_{y_V}^{(k)}$. K is selected depending on the chosen Landmark model (5 for MLD5 and 68 for MLD68).

4. Experiment

In this section, our proposed self-alignment method is evaluated with three popular datasets, including CASIA NIR-VIS 2.0 [3], Oulu-CASIA NIR-VIS [11] and BUAA VIS-NIR [12]. First of all, we introduce these three datasets with training and testing protocols. Secondly, the experiment details are illustrated. Then, the qualitative special alignment and quantitative experimental results are given. Finally, an ablation study is conducted to demonstrate the effect of our proposed methods.

4.1. Datasets and Protocols

The CASIA NIR-VIS 2.0 dataset [3] has the largest number of NIR-VIS images (from 725 subjects), each of which includes 5–50 NIR images and 1–22 VIS images. The images have the same resolution of 640 × 480 but are varied with different properties such as expressions, poses, lighting conditions and whether glasses are worn or not. The Oulu-CASIA NIR-VIS dataset [11] consists of 80 subjects with 6 expressions, covering anger, happiness, sadness, surprise, disgust and fear, as well as three illuminations including darkness, normal indoor lighting and weak light. The BUAA VIS-NIR dataset [12] contains 150 subjects, each of which includes 9 NIR images and 14 VIS images varying in illumination, with diversity in terms of different poses and expressions. We followed the protocols of [3] to split the CASIA NIR-VIS 2.0 dataset into training and testing sets, which contained a total of 10-fold experimental settings. We chose the Rank-1 accuracy, verification rate (VR) at the false accept rate (FAR) = 1%, and VR@FAR = 0.1% for quantitation comparisons on all the three datasets. We used our proposed mean landmark distance (MLD) to test the alignment effect between the generated NIR and VIS images with the same identity.

Following [9], the training sets of the Oulu-CASIA NIR-VIS [11] and the BUAA VIS-NIR [12] datasets were not used. We directly employed our model, trained on the tenth fold of the CASIA NIR-VIS 2.0 dataset for generation, and evaluated it on the testing sets in all three databases as in [40].

4.2. Implementation Details

Our proposed network was trained on the CASIA NIR-VIS 2.0 dataset with two NVIDIA Titan XP GPUs. All images in the dataset were aligned and cropped to 128 × 128 resolutions. For the self-aligned generation part, Adam was used as the optimizer, and the learning rate was fixed to 0.0002. α_1, α_2, α_3, α_4 and α_5 in Equation (9) were set to 50, 5, 1000, 0.2 and 0.001, respectively. We chose a LightCNN [31] model as the ID-feature extractor F_{id}, which was pre-trained on the MS-Celeb-1M database [41]. For the NIR-VIS face recognition part, LightCNN-v29 [31] was selected. We first randomly produced 100,000 paired NIR-VIS images by our well-trained generative model as expanded images, which would boost the limited raw data. Then, we used both the raw data and the generated images to train our recognition network. Stochastic gradient descent (SGD) was adopted as the optimizer, where the momentum was set to 0.9 and the weight decay was set to 5×10^{-4}. The learning rate was set to 7×10^{-4} initially and decayed 1/10 per 5 epochs. The batch size was set to 128, and the dropout ratio was 0.5.

4.3. Alignment Analysis

Owing to our proposed self-aligned encoder–decoder architecture, we could not only generate abundant NIR-VIS images, but also align paired images in the raw dataset with our well-trained model by reconstructing them. Figures 6 and 7 show the reconstructed and generated results of the baseline (DVG), as well as our methods, respectively. Figure 6 is the reconstructed results of DVG and SADG. We divided Figure 7 into two parts; part a is the generated results of DVG and part b shows those of SADG.

Figure 6. Reconstructed images of dual variational generation (DVG) (baseline) and our self-aligned dual generation (SADG) method. For each person, the first column is raw data, the second column is the result of DVG and the third column is the result of SADG. Our results were well-aligned.

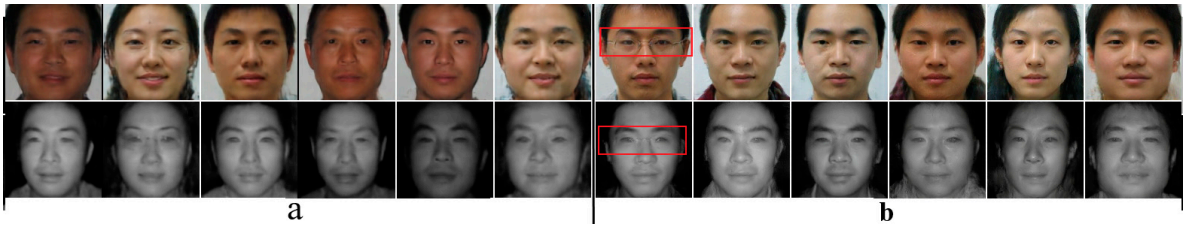

Figure 7. Generated images of DVG (**a**) and SADG (**b**).

It can be observed that DVG could reverse images in the dataset well, yet the reconstructed NIR-VIS images were semantically unaligned, just like the raw data. As shown in Figure 6, the pairwise reconstructed images had different attributes from each other, such as distinct poses, expressions and a pair of glasses. For example, we can see that the VIS image of the first person has no glasses, while the NIR image has glasses. With the devoted reversal process during the training stage, DVG learned two misaligned data distributions according to two modalities and generated misaligned fake images naturally, which is presented in Figure 7a.

Different from DVG, our model could redress the misalignment in the raw dataset and generate paired, well-aligned NIR-VIS images. As shown in Figure 6, glasses were

removed in the reversal NIR image of the first person, and the misaligned poses were redressed in the reversal NIR images of the other person. We can clearly observe that our generated fake images were well-aligned in Figure 7b. It is worth mentioning that even less-obvious glasses were expressed in the paired fake NIR-VIS images, as marked in the first pair of Figure 7b. We further utilized our proposed MLD5 and MLD68 models to evaluate the alignment performance quantitatively.

Table 1 presents the MLD5 and MLD68 models, computed with data in the CASIA NIR-VIS 2.0 dataset, DVG-generated images and the generated images of our model. The MLD5 and MLD68 models of DVG (4.7 and 6.2 pix value per keypoint) were close to those of the raw data, which were 5.2 and 6.5, respectively. This verifies that DVG learned the misaligned data distributions of the two domains from the dataset. The slight improvement of DVG (0.5 in MLD5 and 0.3 in MLD68) may have witnessed the aligning operation, which DVG uses in latent space with L_{dist}. However, it is not enough for alignment in appearance. By contrast, we can obviously see that our generated NIR-VIS images had better performance with 1.4 in MLD5 and 3.1 in MLD68, which substantiates our better alignment capability compared with the baseline.

Table 1. Comparisons of mean landmark distance (MLD) with the CASIA NIR-VIS 2.0 dataset, DVG and our results. Lower values are better.

	CASIA 2.0	DVG	Ours
MLD5	5.2	4.7	1.4
MLD68	6.5	6.2	3.1

4.4. Quantitative Comparison

We compared our proposed self-aligned dual generation (SADG) method with the state-of-the-art approaches on three datasets in this part, as recorded in Table 2. Note that DVG is our baseline method.

Table 2. Comparison of Rank1 accuracy (%) and verification rate (%) on the CASIA NIR-VIS 2.0, Oulu-CASIA NIR-VIS and BUAA VIS-NIR databases.

	CASIA 2.0			Oulu			BUAA		
Method	**Rank1**	**FAR = 1%**	**FAR = 0.1%**	**Rank1**	**FAR = 1%**	**FAR = 0.1%**	**Rank1**	**FAR = 1%**	**FAR = 0.1%**
VGG [40]	62.1	71.0	39.7	-	-	-	-	-	-
LightCNN [31]	96.8	99.1	94.7	96.7	92.4	65.1	96.5	95.4	86.7
PCFH [27]	98.5	99.6	97.3	100	97.7	86.6	98.4	97.9	92.4
PACH [9]	99.0	99.6	98.5	100	97.9	88.2	98.6	98	93.5
DVG [8]	99.9	99.9	99.4	100	98.5	92.9	99.9	98.5	96.6
SADG (Ours)	99.9	99.9	99.6	99.9	98.9	93.2	99.9	98.4	97.3

As for the CASIA NIR-VIS 2.0 dataset, our proposed self-aligned method slightly improved the VR@FAR = 0.1% of the baseline from 99.4% to 99.6% while maintaining the performance of the Rank1 accuracy and VR@FAR = 1%. Compared with the baseline, our method boosted the performance of VR@FAR = 0.1% and VR@FAR = 1% by 0.4% and 0.3%, respectively, on Oulu. Moreover, our model reached 97.3% in VR@FAR = 0.1% in the BUAA dataset, higher than the baseline model by 0.7%. However, the performance of our model slightly droppoed by 0.1% in Rank1 on Oulu and VR@FAR = 1% in the BUAA dataset, compared with the strong baseline (DVG).

On the whole, our method impressively gained improvement compared with the VGG [40] and Light CNN [31] databases, the PACH [27] and PCFH [9] models and the baseline. This demonstrates the better performance of our proposed method. Numerous better-aligned generated NIR-VIS images can tremendously benefit the performance of heterogeneous face recognition network.

4.5. Ablation Study

In this section, we verify the effectiveness of the four segments that were used in our proposed self-aligned generation architecture, including the same z, self-aligned block, slight-weight reconstruction loss and multiscale patch discriminator. Concretely, we set the four contrast experiments as a, b, c and d. Experiment a used a different z code for training in our method, b used the same z code but without our self-aligned block, c used a normal weighted reconstruction loss when training the network and d trained the network without a multiscale patch discriminator.

As for experiment a, when using two different z codes for training, the generated NIR-VIS images showed slight misalignment. As exhibited in Figure 8a, the NIR image of the second person had a misaligned eyebrow with that of the VIS image. In addition, the third paired NIR-VIS images had slightly different poses with each other. The MLD68 in experiment a increased to 4.7 (lower than 6.2 in DVG and higher than 3.1 in ours), and the Rank1 accuracy decreased to 97.2% (lower than 99.9% in ours). This clearly verifies the fact that different z codes can affect the alignment performance virtually, and it can help to bring better alignment and recognition performance with the same latent code z.

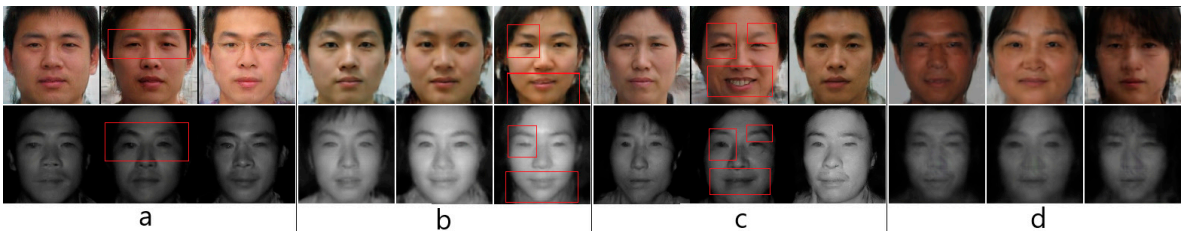

Figure 8. Generated results for the ablation study with contrast experiments (**a**–**d**). (**a**) The same z code. (**b**) Without the self-aligned block. (**c**) Normal weighted reconstruction loss. (**d**) Without a multiscale patch discriminator.

As tabulated in Table 3, b gained a lower MLD68 (3.5) and better recognition accuracy (99.2%) than a, which demonstrates a more important role that the same z value played in the alignment task. That is to say, the same z affected the alignment performance virtually, while the self-aligned block acted the part for redressing the unaligned attributes subsidiarily.

Table 3. Results of the ablation study. a: the same z value; b: without the self-aligned block; c: normal weighted reconstruction loss; d: without a multiscale patch discriminator.

	Same z	Self-Aligned Block	Low Weighted Reconstruction Loss	Multiscale Patch Discriminator	MLD68	Rank1 (%)
Baseline	×	×	×	×	6.2	99.9
a	×	√	√	√	4.7	97.2
b	√	×	√	√	3.5	99.2
c	√	√	×	√	5.8	96.5
d	√	√	√	×	4.1	99.4
SADG	√	√	√	√	3.1	99.9

As for experiment c, we can find the worst results in both appearance and performance in Figure 8c and Table 3. The MLD68 of c reached 5.8, which was nearly close to that of DVG (6.2). Images in Figure 8c show diverse attributes such as the pose, expression and eyebrows. We owe the poor results to the normal weighted reconstruction loss that could close the pix-level distance between the raw data and the reconstructed data by a strong constraint in the training stage. Thus, the results were unaligned, as was the raw data.

When removing the multiscale patch discriminator in experiment d, the generated images were quite bleary, as shown in Figure 8d, and the MLD68 of d increased to 4.1 by 1 pix value per keypoint due to the challenge of detection with blurry images. However,

blurry images slightly affected the recognition performance, because the heterogeneous recognition network was insensitive toward the quality of the images, which was different from human beings. The Rank1 accuracy of d only decreased to 99.4% by 0.5%, as shown in Table 3.

5. Conclusions

In this paper, we first analyzed the misalignment problem in the baseline model. Then, we proposed a self-aligned dual generation (SADG) architecture to cover it, including the self-aligned block and the multiscale patch discriminator. After training our model, we could generate numerous, well-aligned paired NIR-VIS images. We further proposed the mean landmark distance (MLD) for evaluating the alignment performance quantitatively. Extensive experiments on three popular NIR-VIS datasets were conducted, achieving start-of-the-art quantitative results and showing the best alignment performance with our generated images. Finally, an ablation study was performed to demonstrate the effectiveness of our proposed self-aligned generation architecture.

There is still room for the improvement in the field of generating heterogeneous face images. Though we produced numerous semantic, aligned paired NIR-VIS images, we have not improved the attribute diversity of the heterogeneous face data. Semantic face editing shows its ability to produce versatile semantic classes which are nonexistent in the training data. In the future, we may try to utilize that on edited heterogeneous face images based on SADG, boosting the attribute diversity.

Author Contributions: Project administration, P.Z.; validation, P.Z.; investigation, P.Z., F.Z. and Y.Z.; resources, F.Z., X.W. and J.W.; visualization, P.Z., F.Z. and Y.Z.; writing—review and editing, P.Z. and F.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Research and Application of Online-monitoring and Intelligent Emergency Rescue Technology in Hazardous Chemicals Industrial Zone under Grant 19DZ1202200 and in part by the Opening Project of Shanghai Trusted Industrial Control Platform under Grant TICPSH202003004-ZC.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Wang, Z.; Zhang, X.; Yu, P.; Duan, W.; Zhu, D.; Cao, N. A New Face Recognition Method for Intelligent Security. *Appl. Sci.* **2020**, *10*, 852. [CrossRef]
2. Bhatt, H.S.; Bharadwaj, S.; Singh, R.; Vatsa, M. Memetic Approach for Matching Sketches with Digital Face Images. 2012. Available online: https://repository.iiitd.edu.in/jspui/handle/123456789/27 (accessed on 26 March 2012).
3. Li, S.; Yi, D.; Lei, Z.; Liao, S. The casia nir-vis 2.0 face database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 348–353.
4. Liu, S.; Yi, D.; Lei, Z.; Li, S.Z. Heterogeneous face image matching using multi-scale features. In Proceedings of the 2012 5th IAPR International Conference on Biometrics (ICB), New Delhi, India, 29 March–1 April 2012; pp. 79–84.
5. Xue, X.; Han, Z.; Tong, W.; Li, M.; Liu, L. BFRVSR: A Bidirectional Frame Recurrent Method for Video Super-Resolution. *Appl. Sci.* **2020**, *10*, 8749. [CrossRef]
6. Yi, D.; Liu, R.; Chu, R.F.; Lei, Z.; Li, S.Z. Face matching between near infrared and visible light images. In *International Conference on Biometrics (ICB)*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 523–530.
7. Di, X.; Riggan, B.S.; Hu, S.; Short, N.J.; Patel, V.M. Polarimetric Thermal to Visible Face Verification via Self-Attention Guided Synthesis. In Proceedings of the 2019 International Conference on Biometrics (ICB), Crete, Greece, 4–7 June 2019; pp. 1–8.
8. Fu, C.; Wu, X.; Hu, Y.; Huang, H.; He, R. Dual Variational Generation for Low Shot Heterogeneous Face Recognition. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 2674–2683.
9. Yu, J.; Cao, J.; Li, Y.; Jia, X.; He, R. Pose-preserving Cross Spectral Face Hallucination. In Proceedings of the IJCAI, Macao, China, 10–16 August 2019; pp. 1018–1024.
10. Ruan, S.; Tang, C.; Zhou, X.; Jin, Z.; Chen, S.; Wen, H.; Liu, H.; Tang, D. Multi-Pose Face Recognition Based on Deep Learning in Unconstrained Scene. *Appl. Sci.* **2020**, *10*, 4669. [CrossRef]

11. Chen, J.; Yi, D.; Yang, J.; Zhao, G.; Li, S.Z.; Pietikainen, M. Learning mappings for face synthesis from near infrared to visual light images. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Los Angeles, CA, USA, 16–19 June 2019; pp. 156–163.
12. Huang, D.; Sun, J.; Wang, Y. *The Buaa-Visnir Face Database Instructions*; Technology Report IRIP-TR-12-FR-001; School of Computer Science and Engineering, Beihang University: Beijing, China, 2012.
13. Sarfraz, M.S.; Stiefelhagen, R. Deep perceptual mapping for cross-modal face recognition. *Int. J. Comput. Vis. (IJCV)* **2017**, *122*, 426–438. [CrossRef]
14. He, R.; Wu, X.; Sun, Z.; Tan, T. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1761–1773. [CrossRef] [PubMed]
15. Klare, B.; Jain, A.K. Heterogeneous face recognition: Matching nir to visible light images. In Proceedings of the 2010 20th International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, 23–26 August 2010; pp. 1513–1516.
16. Lin, D.; Tang, X. Inter-modality face recognition. In *European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 13–26.
17. Kan, M.; Shan, S.; Zhang, H.; Lao, S.; Chen, X. Multi-View Discriminant Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 188–194. [CrossRef] [PubMed]
18. Huo, J.; Gao, Y.; Shi, Y.; Yang, W.; Yin, H. Heterogeneous face recognition by margin-based cross-modality metric learning. *IEEE Trans. Cybern.* **2017**, *48*, 1814–1826. [CrossRef] [PubMed]
19. Lei, Z.; Liao, S.; Jain, A.K.; Li, S.Z. Coupled discriminant analysis for heterogeneous face recognition. *IEEE Trans. Inf. Forensics Secur.* **2012**, *7*, 1707–1716. [CrossRef]
20. Jin, Y.; Li, J.; Lang, C.; Ruan, Q. Multi-task clustering ELM for VIS-NIR cross-modal feature learning. *Multidimens. Syst. Signal Process.* **2017**, *28*, 905–920. [CrossRef]
21. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
22. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Montréal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
23. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
24. Riggan, B.S.; Short, N.J.; Hu, S.; Kwon, H. Estimation of visible spectrum faces from polarimetric thermal faces. In Proceedings of the 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), New York, NY, USA, 6–9 September 2016; pp. 1–7.
25. Zhang, H.; Riggan, B.S.; Hu, S.; Short, N.J.; Patel, V.M. Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. *Int. J. Comput. Vis. (IJCV)* **2019**, *127*, 845–862. [CrossRef]
26. Zhang, H.; Patel, V.M.; Riggan, B.S.; Hu, S. Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces. In Proceedings of the 2017 IEEE International Joint Conference on Biometrics (IJCB), Denver, CO, USA, 1–4 October 2017; pp. 100–107.
27. Duan, B.; Fu, C.; Li, Y.; Song, X.; He, R. Cross-Spectral Face Hallucination via Disentangling Independent Factors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, Seattle, WA, USA, 14–19 June 2020; pp. 7930–7938.
28. Wu, L.; Du, L.; Liu, B.; Xu, G.; Ge, Y.; Fu, Y.; Li, J.; Zhou, Y.; Hui, X. Heterogeneous metric learning with content-based regularization for software artifact retrieval. In Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 14–17 December 2014; pp. 610–619.
29. Siena, S.; Boddeti, V.N.; Kumar, B.V.K.V. Maximum-margin coupled mappings for cross-domain matching. In Proceedings of the 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), Arlington, VA, USA, 29 September–2 October 2013; pp. 1–8.
30. Zhen, Y.; Rai, P.; Zha, H.; Carin, L. Cross-modal similarity learning via pairs, preferences, and active supervision. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI), Austin, TX, USA, 25–30 January 2015.
31. 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]
32. Mao, Q.; Lee, H.Y.; Tseng, H.Y.; Ma, S.; Yang, M.H. Mode seeking generative adversarial networks for diverse image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Los Angeles, CA, USA, 16–19 June 2019; pp. 1429–1437.
33. Shen, Y.; Gu, J.; Tang, X.; Zhou, B. Interpreting the latent space of gans for semantic face editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, Seattle, WA, USA, 14–19 June 2020; pp. 9243–9252.
34. Shaham, T.R.; Dekel, T.; Michaeli, T. Singan: Learning a generative model from a single natural image. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 4570–4580.
35. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
36. Bao, J.; Chen, D.; Wen, F.; Li, H.; Hua, G. Towards open-set identity preserving face synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6713–6722.

37. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii, HI, USA, 21–26 July 2017; pp. 1125–1134.
38. Sun, Y.; Wang, X.; Tang, X. Deep convolutional network cascade for facial point detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 3476–3483.
39. Zhou, E.; Fan, H.; Cao, Z.; Jiang, Y.; Yin, Q. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV), Darling Harbour, Sydney, Australia, 1–8 December 2013; pp. 386–391.
40. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
41. He, R.; Wu, X.; Sun, Z.; Tan, T. Learning invariant deep representation for nir-vis face recognition. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI), San Francisco, CA, USA, 4–9 February 2017.