Article

Multimodal Human Recognition in Significantly Low Illumination Environment Using Modified EnlightenGAN

Ja Hyung Koo, Se Woon Cho, Na Rae Baek and Kang Ryoung Park *

Division of Electronics and Electrical Engineering, Dongguk University, 30 Pildong-ro 1-gil, Jung-gu, Seoul 04620, Korea; koo6190@dongguk.edu (J.H.K.); jsu319@dongguk.edu (S.W.C.); naris27@dgu.ac.kr (N.R.B.)

* Correspondence: parkgr@dongguk.edu; Tel.: +82-10-3111-7022; Fax: +82-2-2277-8735

Abstract: Human recognition in indoor environments occurs both during the day and at night. During the day, human recognition encounters performance degradation owing to a blur generated when a camera captures a person's image. However, when images are captured at night with a camera, it is difficult to obtain perfect images of a person without light, and the input images are very noisy owing to the properties of camera sensors in low-illumination environments. Studies have been conducted in the past on face recognition in low-illumination environments; however, there is lack of research on face- and body-based human recognition in very low illumination environments. To solve these problems, this study proposes a modified enlighten generative adversarial network (modified EnlightenGAN) in which a very low illumination image is converted to a normal illumination image, and the matching scores of deep convolutional neural network (CNN) features of the face and body in the converted image are combined with a score-level fusion for recognition. The two types of databases used in this study are the Dongguk face and body database version 3 (DFB-DB3) and the ChokePoint open dataset. The results of the experiment conducted using the two databases show that the human verification accuracy (equal error rate (ERR)) and identification accuracy (rank 1 genuine acceptance rate (GAR)) of the proposed method were 7.291% and 92.67% for DFB-DB3 and 10.59% and 87.78% for the ChokePoint dataset, respectively. Accordingly, the performance of the proposed method was better than the previous methods.

Keywords: multimodal human recognition; very low illumination environment; image enhancement; modified EnlightenGAN; CNN

Citation: Koo, J.H.; Cho, S.W.; Baek, N.R.; Park, K.R. Multimodal Human Recognition in Significantly Low Illumination Environment Using Modified EnlightenGAN. *Mathematics* **2021**, *9*, 1934. <https://doi.org/10.3390/math9161934>

Academic Editors: Akemi Galvez Tomida and Bo-Hao Chen

Received: 24 June 2021

Accepted: 9 August 2021

Published: 13 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Various human recognition methods, including face, iris, fingerprint, finger vein, and gait recognition, are continuously being researched. There are limitations to achieving long-distance human recognition in indoor and outdoor environments because recognition becomes increasingly difficult as the human face and iris size decrease with increasing distance between a camera and a person. In addition, a gait recognition method requires a vast amount of data. Owing to these drawbacks, the human body has been the subject of recognition in long-distance human recognition methods in indoor and outdoor environments. Moreover, human recognition methods in indoor and outdoor environments have discrepancies owing to differences in lighting between day and night. A complete image of a person can be easily obtained during the day because the amount of light is sufficient; however, this is not the case at night owing to the inadequate lighting. Therefore, a near-infrared (NIR) camera is also used at night to enable human recognition. However, the number of computations increases as two types of images are captured by both a visible-light camera and an NIR camera. In addition, NIR lighting is required; however, it is difficult to adaptively adjust the NIR light intensity and incidence angle according to the distance between the camera and the object. Considering these problems, this study newly proposes a modified enlighten generative adversarial network (modified EnlightenGAN)

in which a very low illumination image is converted to a normal illumination image, and the matching scores of deep convolutional neural network (CNN) features of the face and body in the converted image are combined with a score-level fusion for recognition. Compared with previous works, our research is novel in the following three ways:

- It is the first study of face- and body-based human recognition in very low illumination images. Accordingly, a modified EnlightenGAN is newly proposed.
- A modified EnlightenGAN is a model that has an improved size (40×40) of the input patch of a discriminator compared to that (32×32) of the conventional EnlightenGAN. In addition, it uses the improved features of the rectified linear unit (ReLU) 5-3 layer in the discriminator for computing the self-feature preserving loss, whereas the conventional EnlightenGAN uses the features of the ReLU 5-1 layer. Therefore, the human recognition performance in a very low illumination image was improved compared to the conventional EnlightenGAN.
- The structural complexity was reduced by separating the modified EnlightenGAN for converting a low-illumination image to a normal illumination one and employing CNNs for human recognition.

2. Related Work

Research on human recognition can be largely divided into recognition that considers low-illumination images and recognition that does not consider low-illumination images. In this section, the previous studies are analyzed based on this classification.

2.1. Recognition That Does Not Consider the Low-Illumination Condition

Many studies have been conducted, to date, on human recognition methods without the low-illumination problem. Human recognition methods can be largely classified as face recognition, body recognition, and face and body recognition. First, for face recognition, Grgic et al. [1] acquired face images from three specified positions applying five cameras. The performance of recognition for the obtained face images was measured using the principal component analysis (PCA) method. Banerjee et al. [2] applied three categories of open databases, ChokePoint, SCface, and FR_SURV, for the experiment, and they measured the performance of recognition using soft-margin learning for multiple feature–kernel combinations (SML-MKFC) with the domain adaptation (DA) method.

Secondly, body recognition can be divided into two types, namely, texture- and shape-based body recognition and gait-based body recognition. For body recognition based on shape and texture, Varior et al. [3] applied the Siamese CNN (S-CNN) structure. Shi et al. [4] used a similar S-CNN architecture in a previous study [3]; however, the researchers used five convolution blocks and a discriminative deep metric learning (DDML) method. However, face recognition and body recognition both have significant drawbacks. The drawbacks of face recognition are that a complete face image can be acquired only if the image is captured from a close distance, and that face recognition becomes difficult in a relatively lower illumination environment. Although body recognition is less affected by the distance than face recognition, it is also very vulnerable to varying illumination conditions. Finally, for gait-based body recognition, Han et al. [5] obtained additional data using a synthetic gait energy image (GEI) method and measured the recognition performance based on the fusion of original GEI and image features obtained through PCA and multiple discriminant analysis (MDA). Gait-based body recognition has the advantage of being less affected by the low-illumination condition; however, it requires a large amount of continuous data, which is a disadvantage.

Furthermore, face and body recognition can be divided into two categories, specifically, face and body recognition based on gait, and body and face recognition based on shape and texture. According to face and body recognition based on gait, Liu et al. [6] computed the performance by applying a database acquired from other research workers, and they used Gabor feature-based elastic bunch graph matching (EBGM) and the hidden Markov model (HMM) as recognition techniques. In a study [7] on texture- and shape-based face

and body recognition, the human face and body were separately recognized in an indoor environment, where the face and body were recognized with visual geometry group (VGG) face net-16 [8] and a residual network (ResNet)-50 [9], respectively.

However, these studies had a poor recognition performance owing to vulnerability to low illumination. The following studies were conducted to overcome the drawbacks of the studies mentioned above.

2.2. Recognition That Considers the Low-Illumination Condition

In general, human recognition needs to consider time-related aspects in addition to recognition methods because a person moves around regardless of the time of day. Accordingly, low illumination is increasingly being considered in research on human recognition. The following studies were conducted on face recognition.

Kamenetsky et al. [10] studied improvements in human recognition in a low-illumination environment as well as recognition methods with respect to distance. An atmospheric turbulence mitigation algorithm (MPE) was used as an improvement method in which MPE was applied for long distances using the reverse heat equation (MPEh), while the MPE based on the median filter (MPEf and fMPE) was applied for a low-illumination environment. The data were obtained 5 m away from a dark indoor room to obtain low-illumination environment data in three different environments of full moon, moonless clear night sky, and moonless overcast. Huang et al. [11] researched recognition in a low-illumination environment using the CASIA-WebFace dataset. Each face image was denoised using a dynamic histogram equalization (DHE), and the recognition performance was improved in a low-illumination environment using a feature reconstruction network (FRN). Poon et al. [12] studied recognition improvement with respect to low illumination and color for which the Asian face database and Faces04 face database were used for low illumination and color, respectively. The PCA method was applied for recognition performance improvement. Zhang et al. [13] used the gradientfaces method to improve the recognition performance in indoor and outdoor environments with varying degrees of illumination. Zhao et al. [14] used the Yale B dataset in which a large area having a low illumination was normalized through the discrete cosine transform (DCT), and the recognition performance was measured by removing a small area of low illumination using a local normalized method. Vu et al. [15] improved the low-illumination environment using two adaptive nonlinear functions and a difference of Gaussian (DoG) filter; Kang et al. [16] performed low-light image denoising using the denoising of low-frequency noise (DeLFN) method; Ren et al. [17] increased the image brightness using a Gaussian homomorphic filter and image multiplication.

However, the application of these methods for face recognition in low-illumination conditions mostly reduced the low-illumination condition without considering the very low illumination condition. Therefore, this study proposes a face and body recognition method that uses the modified EnlightenGAN even in a very low illumination condition. Table 1 summarizes this study and previous studies on low-illumination multimodal human recognition.

Table 1. Overview of the previous studies and the proposed method about low-illumination multimodal human recognition.

Type		Techniques	Strength	Weakness
Not considering low-illumination condition	Face recognition	PCA [1] SML-MKFC with DA [2]	Good recognition performance when images are captured up close	Recognition performance may be degraded owing to external light
	Texture- and shape-based body recognition	S-CNN [3] CNN+DDML [4]	Requires less data for recognition compared to gait-based recognition	Recognition performance may be degraded in a low-illumination environment
	Gait-based body recognition	Synthetic GEL, PCA + MDA [5]	Recognition performance less affected by low illumination	Requires an extended period of time to obtain gait data
	Gait-based body and face recognition	HMM/Gabor feature-based EBGM [6]		
	Body and face recognition based on texture and shape	ResNet-50 and VGG face net-16 [7]	Takes relatively less time to acquire data compared to gait-based recognition	Possibility of clothes color being changed owing to lighting changes or losing important facial features
Considering low-illumination condition	Face recognition	MPEf and fMPE [10]	Able to perform face recognition in a low-illumination condition	Did not consider face and body recognition in a very low illumination environment
		FRN [11]		
		Gradientfaces [12,13]		
		DCT and local normalized method [14]		
		DoG filter and adaptive nonlinear function [15]		
		DeLFN [16]		
	Face and body recognition	Homomorphic filter and image multiplication [17] Proposed method	Recognition possible in a very low illumination environment	Requires more time to process face and body recognition data

3. Proposed Methods

3.1. System Overview

The overall structure of the proposed system is shown in Figure 1. The face region is detected in the input image acquired from an indoor environment (step (1) in Figure 1). The body region below the face region acquired in the previous step is detected (step (2) in Figure 1). In detail, the face region in the captured image is located by the adaptive boosting (AdaBoost) detector [18], and the dlib facial feature tracker [19] is applied to detect the positions of facial features (both eyes), which can more accurately define the face region. Then, the body region is anthropometrically defined based on the size and position of the detected face region. The trained modified EnlightenGAN is used to convert very low illumination in face and body regions into normal illumination (step (3) in Figure 1). Image features are obtained using a CNN model from face and body regions that have been converted to normal illumination (steps (4) and (5) in Figure 1). Imposter and authentic matching distances are measured applying the feature vectors acquired in steps (4) and (5) (steps (6) and (7) in Figure 1). Human recognition is performed (step (9) in Figure 1) using the final matching score obtained through the score-level fusion (step (8) in Figure 1) of two matching distances.

Figure 1. Overall procedure of proposed method.

3.2. Structure of Modified EnlightenGAN

Before introducing the modified EnlightenGAN that is newly proposed in this study, the conventional GAN is briefly explained. A generative adversarial network (GAN)

typically consists of a generator and a discriminator [20]. The generator seeks to produce a counterfeit image resembling the genuine image using Gaussian random noise as an input, while the discriminator aims to discriminate a real image from the fake image generated by the generator. Hence, we train the discriminator to discern a genuine image and a counterfeit image; however, we train the generator to make it certain that the counterfeit image is very similar to the genuine image. Various types of GAN have been proposed owing to their excellent performance. Although DeblurGAN [21], which is a type of conditional GAN (CGAN), generates a fake image in the same manner as a CGAN, it has a very different structure. The generator of a DeblurGAN has two convolutional blocks, nine residual blocks, and two transposed convolution blocks, while the discriminator discerns the input image and target image as concatenated images. Moreover, Pix2pix [22] is a type of CGAN that varies significantly from a conventional CGAN in which skip connection is applied to the encoder-decoder of a generator. Skip connection is applied to solve the problem of a detail feature that occurs when images are processed in the encoder-decoder. Furthermore, a cycle-consistent GAN (CycleGAN) [23] has different characteristics from a CGAN, where there is a relationship between the input image and reference image rather than between the input image and target image in addition to unpaired image-to-image translation. It also varies from a GAN in terms of calculating a loss. If the forward transform output for input X obtained with a CycleGAN is called Y , and then output Y is backward transformed with a CycleGAN to produce the result X' , then the loss for measuring the similarity between X and X' is calculated and is called the cycle consistency loss.

The generator in the EnlightenGAN [24] has the form of a U-net [25]. A U-net is a network that has an encoder and a decoder. The discriminator is further divided into a global discriminator and local discriminator. The global discriminator distinguishes the difference between the real image and fake image generated by the generator, while the local discriminator locally distinguishes two compared images as randomly cropped patch images. The reason for such distinction is that the vanilla discriminator fails to discern lighting changes at a specific position of an image. Both the global and local discriminator thus use a patchGAN. Tables 2–4 present the generator using the discriminator of the modified EnlightenGAN proposed in this study.

Table 2. Generator of modified EnlightenGAN (all the convolution blocks include one convolutional layer, one Leaky rectified linear unit (ReLU) layer, and one batch normalization layer, except for the last convolution block, which includes one convolutional layer and one Leaky ReLU layer).

Type of Layer	Feature Map Sizes (Height $\times$ Width $\times$ Channel)	Number of Filters	Filter Sizes	Number of Strides	Number of Paddings
Input image layer	$224 \times 224 \times 3$				
Attention map layer	$224 \times 224 \times 1$				
Concatenated layer	$224 \times 224 \times 4$				
Convolution block1	$224 \times 224 \times 32$	32	3×3	1×1	1×1
Convolution block2	$224 \times 224 \times 32$	32	3×3	1×1	1×1
Maxpooling layer1	$112 \times 112 \times 32$		2×2	2×2	
Convolution block3	$112 \times 112 \times 64$	64	3×3	1×1	1×1
Convolution block4	$112 \times 112 \times 64$	64	3×3	1×1	1×1
Maxpooling layer2	$56 \times 56 \times 64$		2×2	2×2	
Convolution block5	$56 \times 56 \times 128$	128	3×3	1×1	1×1
Convolution block6	$56 \times 56 \times 128$	128	3×3	1×1	1×1
Maxpooling layer3	$28 \times 28 \times 128$		2×2	2×2	
Convolution block7	$28 \times 28 \times 256$	256	3×3	1×1	1×1

Table 2. Cont.

Type of Layer	Feature Map Sizes (Height $\times$ Width $\times$ Channel)	Number of Filters	Filter Sizes	Number of Strides	Number of Paddings
Convolution block8	$28 \times 28 \times 256$	256	3×3	1×1	1×1
Maxpooling layer4	$14 \times 14 \times 256$		2×2	2×2	
Convolution block9	$14 \times 14 \times 512$	512	3×3	1×1	1×1
Convolution block10	$14 \times 14 \times 512$	512	3×3	1×1	1×1
Deconvolution block1	$28 \times 28 \times 256$	256	3×3	1×1	1×1
Convolution block11	$28 \times 28 \times 256$	256	3×3	1×1	1×1
Convolution block12	$28 \times 28 \times 256$	256	3×3	1×1	1×1
Deconvolution block2	$56 \times 56 \times 128$	128	3×3	1×1	1×1
Convolution block13	$56 \times 56 \times 128$	128	3×3	1×1	1×1
Convolution block14	$56 \times 56 \times 128$	128	3×3	1×1	1×1
Deconvolution block3	$112 \times 112 \times 64$	64	3×3	1×1	1×1
Convolution block15	$112 \times 112 \times 64$	64	3×3	1×1	1×1
Convolution block16	$112 \times 112 \times 64$	64	3×3	1×1	1×1
Deconvolution block4	$224 \times 224 \times 32$	32	3×3	1×1	1×1
Convolution block17	$224 \times 224 \times 32$	32	3×3	1×1	1×1
Convolution block18	$224 \times 224 \times 32$	32	3×3	1×1	1×1
Convolution layer (Output layer)	$224 \times 224 \times 3$	3	1×1	1×1	

Table 3. Global discriminator of modified EnlightenGAN (convolution blocks through 1~6 include one convolutional layer and one Leaky ReLU layer).

Type of Layer	Feature Map Sizes (Height $\times$ Width $\times$ Channel)	Number of Filters	Filter Sizes	Number of Strides	Number of Paddings
Input image layer	$224 \times 224 \times 3$				
Target image layer	$224 \times 224 \times 3$				
Convolution block1	$113 \times 113 \times 64$	64	4×4	2×2	2×2
Convolution block2	$58 \times 58 \times 128$	128	4×4	2×2	2×2
Convolution block3	$30 \times 30 \times 256$	256	4×4	2×2	2×2
Convolution block4	$16 \times 16 \times 512$	512	4×4	2×2	2×2
Convolution block5	$9 \times 9 \times 512$	512	4×4	2×2	2×2
Convolution block6	$10 \times 10 \times 512$	512	4×4	1×1	2×2
Convolution layer	$11 \times 11 \times 1$	1	4×4	1×1	2×2

Table 4. Local discriminator of modified EnlightenGAN (convolution blocks through 1~5 include one convolutional layer and one Leaky ReLU layer).

Type of Layer	Feature Map Sizes (Height $\times$ Width $\times$ Channel)	Number of Filters	Filter Sizes	Number of Strides	Number of Paddings
Layer of input image	$40 \times 40 \times 3$				
Layer of target image	$40 \times 40 \times 3$				
Convolution block1	$21 \times 21 \times 64$	64	4×4	2×2	2×2
Convolution block2	$12 \times 12 \times 128$	128	4×4	2×2	2×2
Convolution block3	$7 \times 7 \times 256$	256	4×4	2×2	2×2
Convolution block4	$5 \times 5 \times 256$	512	4×4	2×2	2×2
Convolution block5	$6 \times 6 \times 512$	512	4×4	1×1	2×2
Convolution layer	$7 \times 7 \times 1$	1	4×4	1×1	2×2

Each convolution block contains a Leaky ReLU layer and a batch-normalized layer [26], as shown in Table 2. If the function of an existing ReLU is $\max(0, x)$, then the function of a Leaky ReLU is $\max(\beta x, x)$, where β is the measure of the slope. The issue with the existing ReLU is that certain neurons output only 0 in the trained model, thus generating a dead ReLU, which is not activated; Leaky ReLU can solve the problem of the dead ReLU as the slope β enables an input below 0 to have an output of multiples of β , instead of simply 0. Figure 2 illustrates the architecture of the modified EnlightenGAN proposed in this study. Figure 2a shows the generator, while Figure 2b shows the discriminator. One of the prominent characteristics of the generator is that an attention map is concatenated with the existing input image (I). An input image is normalized between [0,1] and is calculated as $(1-I)$ to generate an attention map image. An attention module refers to element-wise multiplication between the output of two convolution blocks, and the image is generated through the downsampling of an attention map. Therefore, the attention module at the very bottom has undergone the downsampling layer four times, while the attention module at the very top refers to the attention map image that has not undergone downsampling. The attention map obtained through the attention module is multiplied by the feature map obtained through the convolution layer, thus playing the role of self-regularization [24].

The main differences between the proposed modified EnlightenGAN and the original EnlightenGAN are as follows. A modified EnlightenGAN is a model that has an improved size (40×40) of the input patch of a discriminator compared to that (32×32) of the conventional EnlightenGAN. In addition, it uses the improved features of the ReLU 5-3 layer in the discriminator for computing the self-feature preserving loss of Equation (2), whereas the conventional EnlightenGAN uses the features of the ReLU 5-1 layer.

Figure 2. Cont.

Figure 2. Architecture of modified EnlightenGAN: (a) generator and (b) discriminator.

3.3. Loss Function of Modified EnlightenGAN

The loss function of the modified EnlightenGAN uses self-feature preserving (SFP) loss and original least square GAN (LSGAN) [27] loss [24]. SFP loss performs the role of constraining the VGG feature distance between the low-illumination image and normal illumination image obtained through a generator, emphasizing self-regularization to preserve the image content feature. We can compute the total loss of two loss functions by applying Equation (1) as shown below:

$$L_{total} = L_{SFP} + L_{LSGAN} \quad (1)$$

First, the self-feature preserving loss (L_{SFP}) can be expressed as follows.

$$L_{SFP}(I) = \frac{1}{X_{n,m}Y_{n,m}} \sum_{n=1}^{X_{n,m}} \sum_{m=1}^{Y_{n,m}} (\alpha_{n,m}(I) - \alpha_{n,m}(G(I)))^2 \quad (2)$$

In Equation (2), I denotes the input low-illumination image, and $G(I)$ denotes the generator's enhanced output image. $\alpha_{n,m}$ denotes the feature map extracted from a pre-trained VGG face net-16 model. n denotes the n th max pooling layer, and m denotes the m th convolutional layer after the n th max pooling layer. $X_{n,m}, Y_{n,m}$ are the dimensions for the extracted feature maps. By performing experiments with training data, we chose the optimal parameters of $n = 5$ and $m = 3$ in the modified EnlightenGAN. Each type of loss was applied because the discriminator was divided into global and local for SFP loss. Equations (3) and (4) show the standard functions of a relativistic discriminator.

$$D_R(x_a, x_b) = \delta(B(x_a) - E_{x_b \sim P_f}[B(x_b)]) \quad (3)$$

$$D_R(x_b, x_a) = \delta(B(x_b) - E_{x_a \sim P_r}[B(x_a)]) \quad (4)$$

D_R denotes the standard function, while x_a and x_b refer to real and fake distributions, respectively. B is the discriminator network, while δ is the sigmoid function. For the global discriminator, LSGAN was applied instead of a general sigmoid function. Equations (5) and (6) below, respectively, show the loss functions of the global discriminator and generator.

$$L_{globalD} = E_{x_a \sim P_r} [(D_R(x_a, x_b) - 1)^2] + E_{x_b \sim P_f} [(D_R(x_b, x_a))^2] \quad (5)$$

$$L_{globalG} = E_{x_b \sim P_f} [(D_R(x_b, x_a) - 1)^2] + E_{x_a \sim P_r} [(D_R(x_a, x_b))^2] \quad (6)$$

In Equations (5) and (6), '1' is the target value of $D_R(x_a, x_b)$ and $D_R(x_b, x_a)$. As shown in Equations (3) and (4), $D_R(x_a, x_b)$ and $D_R(x_b, x_a)$ are the outputs of the sigmoid function of $B(x_a) - E_{x_b \sim P_f} [B(x_b)]$ and $B(x_b) - E_{x_a \sim P_r} [B(x_a)]$, respectively. If the input increases, the output of the sigmoid function converges to 1. Therefore, in Equations (5) and (6), our network is trained to make both $B(x_a) - E_{x_b \sim P_f} [B(x_b)]$ and $B(x_b) - E_{x_a \sim P_r} [B(x_a)]$ larger, which causes both $D_R(x_a, x_b)$ and $D_R(x_b, x_a)$ to converge to '1'.

For the local discriminator, eight patches between the output image and real image in the modified EnlightenGAN were randomly cropped. Equations (7) and (8) below, respectively, show the local discriminator and generator loss applied with the original LSGAN and adversarial loss.

$$L_{localD} = E_{x_a \sim P_{r-patches}} [(D(x_a) - 1)^2] + E_{x_b \sim P_{f-patches}} [(D(x_b) - 0)^2] \quad (7)$$

$$L_{localG} = E_{x_a \sim P_{f-patches}} [(D(x_b) - 1)^2] \quad (8)$$

In Equations (7) and (8), '1' is the target value of the discriminator output of real data ($D(x_a)$), whereas '0' is the target value of the discriminator output of fake data ($D(x_b)$). That is, our network is trained to make $D(x_a)$ close to '1' and $D(x_b)$ close to '0'.

Global and local discriminators were calculated separately for the SFP loss, and Equation (1) can therefore be ultimately expressed as shown in Equation (9).

$$L_{total} = L_{SFP(G)} + L_{SFP(L)} + L_{global(G)} + L_{local(G)} \quad (9)$$

3.4. Deep CNNs and Score-Level Fusion for Face and Body Recognition

Face and body images converted to normal illumination with the modified EnlightenGAN were recognized using VGG face net-16 [8] and ResNet-50 [9]. Although the two models are already pre-trained, finetuning was performed for the training data applied in this research. VGG face net-16, applied for images of the face, is composed of a neural network and convolution filters. Distinctly, it is composed of 13 convolutional layers, 3 fully connected layers, and 5 pooling layers. The pre-trained CNN model which was applied in this research was trained with YouTube faces [28] and Labeled Faces in the Wild [29]. Normal illumination images obtained by the modified EnlightenGAN perform computations using a convolution filter. The equation for the output feature map obtained by a convolution filter is $\text{output} = (M - C + 2D)/G + 1$. M indicates the height and width size of the input, C indicates the convolutional layer filter size, D indicates the number of paddings, and G indicates the number of strides. For instance, if a 256×256 image goes through a convolution filter as $C = 6$, $D = 2$, and $G = 2$, then the output size is $(256 - 6 + 4)/2 + 1$, i.e., 128×128 .

Similar to VGG face net-16, ResNet has a large number of versions depending on the number of layers. As the number of layers increases, the processing speed and important feature loss of the input image increase owing to the convolution computation. To reduce the operation speed of the convolution layer, a bottleneck architecture was proposed for ResNet. An example of a bottleneck is as follows. If two 3×3 convolution layers are converted to 1×1 , 3×3 , and 1×1 convolutional layers, the computation speed can be reduced while producing the same results. The 1×1 convolutional layer in the front is for

reducing the dimension, while the 1×1 convolutional layer in the back is for expanding the dimension again. The feature loss of ResNet is a vanishing or exploding gradient problem, and ResNet uses a shortcut to solve this problem. If the input passes through two convolutional layers, the convolution is computed two times. If the result of computing two convolutional layers is called $F(x)$, the sum of the short structure can be expressed as the sum of $F(x)$ and input X , or $F(x) + X$. The shortcut aims to reduce the feature loss by adding a previous input feature.

Based on 4096 features acquired from a fully connected (FC) layer of VGG face net-16 and 2048 features obtained from the average pooling layer of ResNet-50, two matching scores were, respectively, calculated by the Euclidean distance with the registered 4096 and 2048 features. These two scores had different scales and thus needed to be converted to the same numerical range. Min–max normalization was performed for the maximum and minimum scores extracted based on the training data using each recognition method to convert the score range to a value between 0 and 1. As shown in step (8) in Figure 1, these two matching scores were combined through score-level fusion to obtain the final matching score. Recognition performance was compared by applying the weighted sum, weighted product [30], and support vector machine (SVM)-based [31] score-level fusion in Equations (10)–(13) to the two normalized matching scores. Through experiments with training data, the optimal kernel of the radial basis function (RBF) was selected in the SVM.

$$S_{ws} = w \times S_1 + (1 - w) \times S_2 \quad (10)$$

$$S_{wp} = S_1^w \times S_2^{1-w} \quad (11)$$

$$S_{svm} = \text{sign}(\sum a_i y_i K(S_i, S_j) + b) \quad (12)$$

$$\text{RBF kernel : } K(S_i, S_j) = e^{-\gamma S_i - S_j^2} \gamma > 0 \quad (13)$$

Here, S_1 and S_2 refer to the scores of VGG face net-16 and ResNet-50, respectively. In addition, w refers to the weight. a_i , y_i , and b are the trained parameters [31]. Optimal weights in Equations (10)–(13) were obtained to minimize the error (EER) of recognition in the training data. In detail, the EER of recognition was measured with training data according to various weights w , parameters, and gammas. From this procedure, the optimal weights w , parameters, and gammas were determined, which caused the minimal EER of recognition. The following method was used to determine genuine matching or imposter matching based on the final matching score obtained through score-level fusion. Specifically, it is considered genuine matching if the score is lower than the threshold, or imposter matching if the score is higher than the threshold determined with respect to the EER of the genuine matching and imposter matching distribution acquired from the training data. Genuine matching refers to the case when the registered image and recognition image have the same class, while imposter matching refers to the case when the registered image and recognition image have different classes. Moreover, EER is the error rate when the false acceptance rate (FAR), which is the error rate of incorrectly accepting imposter data as genuine, becomes the same as the false rejection rate (FRR), which is the error rate of incorrectly rejecting genuine data as fake.

4. Experimental Results and Analysis

4.1. Experimental Environment and Database

We used two categories of cameras in this research in order to obtain the Dongguk face and body database version 3 (DFB-DB3), namely, the Logitech BCC950 [32] and the Logitech C920 [33]. Figure 3 shows examples of the captured images. Figure 3a shows the images captured with the Logitech BCC950 camera, while Figure 3b shows the images captured with the Logitech C920 camera. Figure 3c shows an example of an image obtained by adjusting the gamma value of 5 only to the intensity value (V) among HSV components of the normal illumination image to convert it into a low-illumination image.

Figure 3. Example of DFB-DB3 figures obtained from (a) the Logitech C920 camera and (b) the Logitech BCC950 camera. (c) Converted low-illumination figure of DFB-DB3.

The ChokePoint dataset, which is an open database, was used in the experiment to evaluate the generality of the proposed method. The ChokePoint dataset [34] is provided free of charge by National ICT Australia Ltd. (NICTA) and is composed of Portals 1 and 2. Portal 1 consists of 25 persons (6 female and 19 male), while Portal 2 consists of 29 persons (6 female and 23 male). The ChokePoint dataset was acquired from six positions using three cameras. Portal 2 satisfied the conditions of DFB-DB3 used in this study; the data of 28 individuals (22 male and 6 female) were used for two-fold cross-validation. Similar to DFB-DB3, the gamma value of 5 was adjusted only to the intensity value (V) among HSV components to convert the image into a low-illumination image, as shown in Figure 4.

Table 5 presents the number of images of the DFB-DB3 and the ChokePoint dataset used in this study. Both datasets were assessed with two-fold cross-validation, where each dataset was further split into sub-datasets 1 and 2. For instance, if sub-dataset 2 was applied for training, sub-dataset 1 was applied for testing. If sub-dataset 1 was applied for training, sub-dataset 2 was applied for testing to estimate the recognition execution. Each sub-dataset was configured to include the images of different classes (persons). That is, the persons whose data were used for training were completely different from those

whose data were used for testing. Therefore, the clothes of persons whose data were used for training were also completely different from those whose data were used for testing.

Figure 4. Example images for ChokePoint dataset. (a) Original images of ChokePoint dataset [34]. (b) Converted low-illumination image of ChokePoint dataset.

Table 5. Overall images of ChokePoint dataset and DFB-DB3.

		DFB-DB3			ChokePoint Dataset		
		Number of Each Fold Class	Number of Testing Images	Number of Augmented Training Images	Number of Each Fold Class	Number of Testing Images	Number of Augmented Training Images
Face	Sub-Dataset1	11	827	200,134	14	10,381	332,192
	Sub-Dataset2	11	989	239,338	14	10,269	328,608
Body	Sub-Dataset1	11	827	200,134	14	10,381	332,192
	Sub-Dataset2	11	989	239,338	14	10,269	328,608

The computer specification applied for the experiment is as follows: 16 GB RAM, CPU Intel(R) Core(TM) i7-6700 CPU @ 3.40 GHz, compute unified device architecture (CUDA) version 10.0 [35], and NVIDIA GeForce GTX 1070 graphics processing unit (GPU) card [36]. The EnlightenGAN and CNN were implemented using PyTorch 1.0.1 [37] and Caffe [38], respectively.

4.2. Training of Modified EnlightenGAN and CNN Models

The learning rate and batch size of EnlightenGAN were 0.00001 and 1, respectively. The training epoch was 200. As an optimizer, adaptive moment estimation (Adam) was used [39]. After converting to a normal illumination image with EnlightenGAN, VGG face net-16 [8] was used for the face image, while ResNet-50 [9] was used for the body image. VGG face net-16 is a model pre-trained with the Oxford face database [40]; however, this model was fine-tuned to be suitable for the features of DFB-DB3. Moreover, ResNet-50 is a pre-trained model that was fine-tuned to be suitable for the features of body images applied in this research. We performed training by using a 0.0001 learning rate and a batch size of 20 on VGG face net-16 and a batch size of 15 on ResNet-50. Moreover, the epochs for training were 20 and 15 on VGG face net-16 and ResNet-50, respectively. As an optimizer, stochastic gradient descent (SGD) was used [39].

A sufficient amount of data is generally needed to train a CNN model, and data augmentation was therefore applied. With respect to Table 5, we applied data augmentation just for the data of the training, as shown in Figure 5. Image translation together with image flipping was applied for data augmentation; according to image translation, the image center was cropped and then translated in all four directions. For image flipping, the images were horizontally flipped. The training data of DFB-DB3 were translated by five pixels in all four directions, while the ChokePoint dataset was translated by two pixels in the up and down directions and by one pixel in the right and left directions. Around 440,000 images were acquired from sub-datasets 1 and 2 of the training data of DFB-DB3; around 660,000 images were acquired from sub-datasets 1 and 2 of the ChokePoint dataset. Figure 6 illustrates the CNN model accuracy–loss graph that was trained with body and face images. The training loss and accuracy graphs of all cases converged to 0 as the iterations increased, which indicates that the CNN model used in this study was sufficiently trained.

Figure 5. Method for data augmentation including (a) cropping and image translation, and (b) horizontal flipping.

Figure 6. Cont.

Figure 6. Graphs illustrating training accuracy and loss of DFB-DB3 (a–d) result and ChokePoint dataset (e–h) result. VGG face net-16 with respect to (a,e) the first fold and (b,f) the second fold. ResNet-50 with respect to (c,g) the first fold and (d,h) the second fold.

4.3. Testing of Modified EnlightenGAN and CNN Models with DFB-DB3

4.3.1. Ablation Studies

To evaluate the illumination enhancement performance of the modified EnlightenGAN, the similarity between the original normal illumination image and illumination-enhanced image obtained using the modified EnlightenGAN was measured using the signal-to-noise ratio (SNR) [41], peak signal-to-noise ratio (PSNR) [42], and structural similarity (SSIM) [43], as expressed in Equations (14)–(17).

$$\text{MSE} = \frac{1}{hw} \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} [I_i(i, j) - I_o(i, j)]^2 \quad (14)$$

$$\text{SNR} = 10 \log_{10} \left(\frac{\frac{\sum_{i=0}^{h-1} \sum_{j=0}^{w-1} [I_i(i, j)]^2}{hw}}{\text{MSE}} \right) \quad (15)$$

$$\text{PSNR} = 10 \log_{10} \left(\frac{255^2}{\text{MSE}} \right) \quad (16)$$

I_o is the illumination-enhanced image, whereas I_i is the original normal illumination image. h and w are, respectively, the height and width of the image.

$$\text{SSIM} = \frac{(2\mu_i\mu_o + C_1)(2\sigma_{io} + C_2)}{(\mu_i^2 + \mu_o^2 + C_1)(\sigma_i^2 + \sigma_o^2 + C_2)} \quad (17)$$

μ_o and σ_o are, respectively, the mean and standard deviation of the pixel values of the illumination-enhanced image; μ_i and σ_i are, respectively, the mean and standard deviation of the pixel values of the original normal illumination image. σ_{io} is the covariance of the two images. C_1 and C_2 are constants used to prevent the denominator of each equation from becoming 0.

For the first ablation study, the performance was compared in terms of the number of patches used in the discriminator, as shown in Table 6; the experiment results show that the performance was best when eight patches were used, whereas the original EnlightenGAN uses five patches [24].

In the next ablation study, the performances of the modified EnlightenGAN and original EnlightenGAN were compared. As presented in Table 7 and Figure 7, the original EnlightenGAN shows a slightly better performance than the modified EnlightenGAN in

terms of the quality of the generated image. However, the proposed modified EnlightenGAN outperforms the original EnlightenGAN in terms of recognition accuracies.

Table 6. Performance comparisons with enhanced images by modified EnlightenGAN according to the number of patches in the discriminator.

Method	Number of Patches	Face	Body
SNR	5	15.532	14.172
	8	19.294	15.343
	11	11.237	11.2
PSNR	5	24.42	23.533
	8	28.181	24.704
	11	20.125	20.61
SSIM	5	0.78	0.703
	8	0.896	0.769
	11	0.727	0.697

(a)

(b)

(c)

Figure 7. Cont.

Figure 7. Comparisons of output images by original EnlightenGAN and modified EnlightenGAN. (a) Original normal illumination image. (b) Low-illumination image. Output images by (c) original EnlightenGAN, and (d) modified EnlightenGAN.

Table 7. Performance comparisons with enhanced images by original EnlightenGAN or modified EnlightenGAN.

Method	Original EnlightenGAN [24]		Modified EnlightenGAN	
	Face	Body	Face	Body
SNR	19.59	18.852	19.294	15.343
PSNR	28.478	28.213	28.181	24.704
SSIM	0.895	0.849	0.896	0.769

For the next ablation study, the recognition accuracies (EER) of the face and body were compared for the cases with and without the usage of the modified EnlightenGAN. In addition, the recognition accuracies (EER) obtained by various score-level fusions explained in Section 3.4. were compared for the cases with and without the usage of the modified EnlightenGAN. Tables 8 and 9 show that the recognition accuracies were improved with the modified EnlightenGAN. Proceeding with recognition by converting the image to a normal illumination image using the modified EnlightenGAN had a better effect on the recognition performance than when using a low-illumination image. Figure 8 illustrates the receiver operating characteristic (ROC) curves for the recognition accuracies. Here, the genuine acceptance rate (GAR) is calculated as $100 - \text{FRR} (\%)$. As shown in Figure 8, the performance with the modified EnlightenGAN is better than that without the modified EnlightenGAN. Furthermore, the weighted product method exhibited a better performance than other score-level fusion methods, as shown in Table 9 and Figure 8.

Table 8. Comparisons of EER for face and body recognition with or without modified EnlightenGAN (unit: %).

Method		Face	Body
Without modified EnlightenGAN	1st fold	20.22	22.08
	2nd fold	29.32	21.32
	Average	24.77	21.7
With modified EnlightenGAN	1st fold	11.21	19
	2nd fold	8.44	19.76
	Average	9.825	19.38

Figure 8. ROC curves of recognition accuracies with or without modified EnlightenGAN. Results of (a) face and body recognition, and (b) various score-level fusions.

Table 9. Comparisons of EER for score-level fusion with or without modified EnlightenGAN (unit: %).

Method		Score-Level Fusion		
		SVM	Weighted Product	Weighted Sum
Without modified EnlightenGAN	1st fold	33.62	16.979	17.077
	2nd fold	25.73	21.605	21.345
	Average	29.675	19.292	19.211
With modified EnlightenGAN	1st fold	11.552	8.805	8.848
	2nd fold	11.51	7.735	7.812
	Average	11.531	8.27	8.33

4.3.2. Comparisons of Proposed Method with State-of-the-Art Methods

First, the modified EnlightenGAN and other GAN-based methods were compared to obtain a comparison with state-of-the-art low-illumination image enhancement methods. As shown in Table 10 and Figure 9, the modified EnlightenGAN exhibited the best performance for face images; however, the recognition performance of Pix2pix [22] was the best for body images. One of the advantages of Pix2pix is that the input image and target image are concatenated, thus exhibiting a better performance than EnlightenGAN. Owing to the nature of EnlightenGAN, an output image of an input image may not have a good quality depending on the state of the attention map. As shown in Table 10 and Figure 9, the best recognition performance was obtained when score-level fusion was applied to the results of the modified EnlightenGAN (face) + Pix2pix (body) based on the weighted product rule.

Table 10. Comparisons of EERs using our proposed method with those using state-of-the-art GAN-based methods (unit: %).

Method		1st Fold	2nd Fold	Average
Face	Modified EnlightenGAN	11.21	8.44	9.825
	Original EnlightenGAN [24]	12.51	8.53	10.52
	CycleGAN [23]	14.2	6.47	10.335
	Pix2pix [22]	12.26	7.7	9.98
Body	Modified EnlightenGAN	19	19.76	19.38
	Original EnlightenGAN [24]	18.52	19.47	18.995
	CycleGAN [23]	21.89	21.38	21.635
	Pix2pix [22]	17.04	18.39	17.715
Weighted sum	Modified EnlightenGAN (face) + Pix2pix (body)	7.819	7.034	7.427
	Modified EnlightenGAN (face and body)	8.848	7.812	8.33
	CycleGAN [23]	12.067	5.866	8.967
	Pix2pix [22]	10.466	5.506	7.986
Weighted product	Modified EnlightenGAN (face) + Pix2pix (body)	7.623	6.958	7.291
	Modified EnlightenGAN (face and body)	8.805	7.735	8.27
	CycleGAN [23]	10.767	5.659	8.213
	Pix2pix [22]	9.35	5.399	7.375
SVM	Modified EnlightenGAN (face) + Pix2pix (body)	9.72	7.12	8.42
	Modified EnlightenGAN (face and body)	11.552	11.51	11.531
	CycleGAN [23]	13.32	6.57	9.945
	Pix2pix [22]	9.08	5.9	7.49

(a)

(b)

Figure 9. Cont.

Figure 9. Graph for ROC curves acquired using our method and the previous GAN-based techniques. (a,b) Recognition results for face and body images, and (c) score-level fusion result.

Secondly, the previous recognition methods and performances for face and face and body were compared. As shown in Tables 11 and 12 and Figure 10, the proposed method had a more outstanding recognition performance than the state-of-the-art methods.

Table 11. Comparison of EERs using the proposed method and the previous face recognition methods (unit: %).

Method	1st Fold	2nd Fold	Average
Proposed method	7.623	6.958	7.291
VGG face net-16 [44]	25.21	35.59	30.4
ResNet-50 [45,46]	22.92	30.13	26.525

Table 12. Comparison of EERs using the proposed method and the previous body and face recognition techniques (unit: %).

Method	1st Fold	2nd Fold	Average
ResNet IDE + LIME [47]	20.3	29.26	24.78
ELF [48]	28.02	30.07	29.045
Proposed method	7.623	6.958	7.291

Figure 10. ROC curves acquired by the previous methods and proposed method. (a) Recognition results for face image, and (b) recognition results for body and face image.

Thirdly, the recognition performance was evaluated based on the cumulative match characteristic (CMC) curve, which is often used for 1:n matching. Figure 11 shows the comparison results of the proposed method and the state-of-the-art methods in Tables 11 and 12. The vertical axis indicates the GAR value for each rank, and the horizontal axis indicates the rank. As the number of classes is 11 in Table 5, the maximum rank is 11, as shown in Figure 11. As shown in Figure 11b, our proposed method (rank 1 GAR of 92.67%) outperforms the state-of-the-art methods in the CMC curves.

Figure 12 illustrates the statistical difference in the result by estimating the t -test results [49] and Cohen's d-value [50] by the proposed method and the second best method. The d-value of ResNet-50 [45,46] and the proposed method on face recognition is 5.31, thus having a significant effect (large effect size) because the value is greater than 0.8, while the p -value of the t -test results is around 0.033, which indicates a significant difference with the proposed method at the confidence level of 95%. Furthermore, for body and face recognition, the t -test results and Cohen's d-value were estimated with respect to ResNet IDE + LIME [47], which has the second best recognition result in comparison with the proposed method, where Cohen's d-value was 3.89 (large effect size); this implies that there was a significant effect, while the t -test results had a significant difference at the confidence level of 94%.

The false acceptance (FA), false rejection (FR), and correct recognition cases obtained by our proposed method are analyzed in Figure 13. Figure 13 shows different examples in which the left-side image denotes the registered image, and the right-side image denotes the inquiry (recognized) image. The right part of the red box image is converted from the proposed method. As shown in Figure 13, FA occurred for similar clothes between different people, and FR occurred for frequent changes in facial features. Correct recognition was

observed by the proposed method even if the face image was severely blurred by low illumination, as shown in Figure 13c.

(a)

(b)

Figure 11. CMC curves of the proposed and state-of-the-art methods. (a) Face recognition results obtained by the proposed and state-of-the-art methods, and (b) face and body recognition results obtained by the proposed and state-of-the-art methods.

(a)

Figure 12. Cont.

Figure 12. Graphs for *t*-test result between the second best model and our proposed method with respect to accuracy of average recognition. (a) Comparison between ResNet-50 and the proposed method, and (b) comparison between ResNet IDE + LIME and the proposed method.

Figure 13. Cont.

Figure 13. Cases of FA, FR, and correction recognition. (a) Cases of FA, (b) cases of FR, and (c) correct cases. In (a–c), left and right images are enrolled and recognized images, respectively.

4.3.3. Class Activation Map

Consequently, the class activation feature maps of ResNet-50 and VGG face net-16 of DFB-DB3 used for measuring the performance of recognition for body and face images were analyzed. Figure 14 illustrates the class activation feature map of a particular layer applying the Grad-CAM method [51]. Significant characteristics are exhibited via a division map. In Figure 14, the first image is the original input image, the second image is the converted low-illumination image, the third image is the input image of face and body images (restored by our modified EnlightenGAN) in the CNN model, and the fourth~sixth images are the class activation feature maps of body and face images.

Figure 14. Results on class activation feature map of DFB-DB3. (a,b) are the activation map results from face images. From the left image to the right image: original image, low-illumination image, enhancement image from modified EnlightenGAN, result from 7th ReLU layer, result from 12th ReLU layer, and result from 13th ReLU layer of VGG face-net 16. (c,d) are the activation map results from body images. From the left image to the right image: original image, low-illumination image, enhancement image from modified EnlightenGAN, result from 3rd batch-normalized layer, result from conv5 2nd block, and result from conv5 3rd block of ResNet-50.

In detail, as the image of the input (Figure 14-3rd) is handled via VGG face net-16, Figure 14-4th denotes the seventh ReLU layer's class activation feature map. Figure 14-5th denotes the 12th ReLU layer's class activation feature map. Figure 14-6th denotes the 13th ReLU layer's class activation feature map. Figure 14-6th represents the focused distribution part near the face region, where the red color shows the important characteristic, whereas the blue color shows less significant features. The black color means that there were no important features. When the procedure undergoes Figure 14-5th through Figure 14-6th, the important features are more concentrated near the face area. In addition, body images were obtained at the batch-normalized layer. Contrary to the results for face images, the important features were represented near the body area since the trained section of the ResNet-50 model regards data that pertain to the person's clothes and body as important features.

4.4. Testing of Modified EnlightenGAN and CNN Models with ChokePoint Dataset

4.4.1. Ablation Studies

In the first experiment conducted using the ChokePoint open database, the recognition accuracies (EER) of the face and body were compared with or without the modified EnlightenGAN. In addition, the recognition accuracies (EER) obtained by various score-level fusions explained in Section 3.4 were compared with and without the modified

EnlightenGAN. Tables 13 and 14 show that the recognition accuracies were improved with EnlightenGAN. Proceeding with recognition by converting to a normal illumination image using the modified EnlightenGAN had a greater effect on the recognition performance than when using a low-illumination image. Figure 15 illustrates the ROC curves for the recognition accuracies. As shown in Figure 15, the performance with the modified EnlightenGAN is better than that without the modified EnlightenGAN. Furthermore, the weighted sum exhibited a better performance than other score-level fusion methods, as shown in Table 14 and Figure 15.

Table 13. Comparisons of EER for face and body recognition with and without modified EnlightenGAN (unit: %).

Method		Face	Body
Without modified EnlightenGAN	1st fold	25.5	38.06
	2nd fold	28.49	32.07
	Average	26.995	35.065
With modified EnlightenGAN	1st fold	13.25	28.96
	2nd fold	11.47	25.94
	Average	12.36	27.45

Table 14. Comparisons of EER for score-level fusion with and without modified EnlightenGAN (unit: %).

Method		Score-Level Fusion		
		Weighted Sum	Weighted Product	SVM
Without modified EnlightenGAN	1st fold	24.44	24.4	40.16
	2nd fold	23.52	23.47	40.02
	Average	23.98	23.935	40.09
With modified EnlightenGAN	1st fold	12.59	12.542	23.82
	2nd fold	19.83	19.928	17.16
	Average	16.21	16.235	20.49

(a)

Figure 15. Cont.

Figure 15. ROC curves of recognition accuracies with and without modified EnlightenGAN. Results of (a) face and body recognition, and (b) various score-level fusions.

4.4.2. Comparing between the Proposed Method and the Previous Techniques

The modified EnlightenGAN and other GAN-based methods were compared to obtain a comparison with state-of-the-art low-illumination image enhancement methods. As shown in Table 15 and Figure 16, the modified EnlightenGAN exhibited the second best performance for face images. However, the recognition performance of Pix2pix [22] was worse for body images. Nevertheless, one of the advantages of Pix2pix is that the input image and target image are concatenated, thus exhibiting a performance that is worse than EnlightenGAN but also exhibiting a resulting image that is better than EnlightenGAN. Owing to the nature of EnlightenGAN, the output image of an input image may not have a good quality depending on the state of the attention map. As shown in Table 15 and Figure 16, the best recognition performance was obtained when score-level fusion was applied to the results of the modified EnlightenGAN (face) + Pix2pix (body) based on the weighted product rule.

Table 15. Comparisons of EERs obtained using our proposed method with those obtained using state-of-the-art GAN-based methods (unit: %).

	Method	1st Fold	2nd Fold	Average
Face	Modified EnlightenGAN	13.25	11.47	12.36
	Original EnlightenGAN [24]	12.88	15.17	14.025
	CycleGAN [23]	15.87	11.11	13.49

Table 15. Cont.

	Method	1st Fold	2nd Fold	Average
Body	Modified EnlightenGAN	28.96	25.94	27.45
	Original EnlightenGAN [24]	25.01	23.86	24.435
	CycleGAN [23]	27.16	29.56	28.36
Weighted sum	Modified EnlightenGAN (face) + Pix2pix (body)	11.821	9.54	10.681
	Modified EnlightenGAN (face and body)	12.59	9.92	11.255
	CycleGAN [23]	13.69	10.32	12.005
	Pix2pix [22]	9.835	13.43	11.633
Weighted product	Modified EnlightenGAN (face) + Pix2pix (body)	11.689	9.49	10.59
	Modified EnlightenGAN (face and body)	12.542	9.87	11.206
	CycleGAN [23]	13.59	10.23	11.91
	Pix2pix [22]	9.843	14.31	12.08
SVM	Modified EnlightenGAN (face) + Pix2pix (body)	17.17	15.51	16.34
	Modified EnlightenGAN (face and body)	23.82	17.16	20.49
	CycleGAN [23]	21.15	16.83	18.99
	Pix2pix [22]	15.86	8.76	12.31

(a)

(b)

Figure 16. Cont.

Figure 16. ROC curves acquired by our proposed method and previous GAN-based methods. (a,b) Recognition results for face and body images, and (c) score-level fusion result.

Secondly, state-of-the-art recognition methods and the performances for the face and face and body were compared. As shown in Tables 16 and 17 and Figure 17, the proposed method had a more outstanding recognition performance than the state-of-the-art methods.

Table 16. Comparisons of EERs obtained using proposed method and state-of-the-art face recognition methods (unit: %).

Method	1st Fold	2nd Fold	Average
Proposed method	11.689	9.49	10.59
VGG face net-16 [44]	32.72	39.61	36.165
ResNet-50 [45,46]	39.94	41.21	40.575

Table 17. Comparisons of EERs obtained using proposed method and state-of-the-art body and face recognition techniques (unit: %).

Method	1st Fold	2nd Fold	Average
ResNet IDE + LIME [47]	30.88	49.47	40.175
ELF [48]	36.27	30.6	33.435
Proposed method	11.689	9.49	10.59

Figure 18 illustrates the CMC curves of the proposed method and state-of-the-art methods for face and face and body recognition. As the number of classes is 14 in Table 5,

the maximum rank is 14, as shown in Figure 18. As shown in Figure 18b, the performance of the proposed method (rank 1 GAR of 87.78%) exceeded that of the state-of-the-art methods.

Figure 17. ROC curves acquired using the proposed method and the previous methods. (a) Results for face recognition, and (b) results for body and face recognition.

Figure 18. CMC curves of the proposed and state-of-the-art methods. (a) Face recognition results obtained by the proposed and state-of-the-art methods, and (b) face and body recognition results obtained by the proposed and state-of-the-art methods.

Figure 19 represents the statistical performance dissimilarity obtained by estimating the t -test and Cohen's d -value results of the proposed method and the second best method. For face recognition, the d -value between the proposed method and VGG face net-16 [44] is 7.07, thus having a significant effect (large effect size) because the value is greater than 0.8, while the p -value of the t -test results is around 0.019, which indicates a significant difference with the proposed method at a confidence level of 95%. Furthermore, for body and face recognition, the t -test and Cohen's d -value results were estimated with respect to ELF [48], which has the second best performance in comparison with the proposed method, where Cohen's d -value was 7.51; this implies that there was a significant effect (large effect size), while the t -test results had a significant difference at a 95% confidence level.

Figure 19. Graphs for t -test result of the second best model and our proposed method with regard to average recognition accuracy. (a) Comparison of VGG face net-16 [44] and the proposed method, and (b) comparison of the proposed method and ELF [48].

The FA, FR, and correct recognition cases obtained by our proposed method are analyzed in Figure 20. Figure 20 shows dissimilar examples, where the left-side image denotes the registered image, and the right-side image denotes the inquiry (recognized) image. The red box part of the image on the right was converted by applying the proposed method. As shown in Figure 20, FA occurred for similar colors of clothes, and FR occurred for the brightness changes of faces. Correct recognition was observed by the proposed method for blurred and different colors of faces, as shown in Figure 20c.

(a)

(b)

Figure 20. Cont.

Figure 20. Cases of FA, FR, and correction recognition on ChokePoint database [34]. (a) Cases of FA, (b) cases of FR, and (c) correct cases. In (a–c), the left and right images are enrolled and recognized images, respectively.

4.4.3. Class Activation Feature Map

In the following experiment, the class activation feature map of the ChokePoint dataset was considered. Figure 21 represents the results of the class activation feature map. The face image relates to the class activation feature map acquired by the ReLU layer in VGG face net-16. In Figure 21, the first image is the original input image, the second image is

the converted low-illumination image, the third image is the input image of face and body images (restored by the modified EnlightenGAN) in the CNN model, and the fourth~sixth images are the class activation feature maps of the face and body images. The fourth~sixth images show that the distribution converged near the face region, in which the red color denotes the main characteristic, whereas the blue color denotes minor features. The black color denotes undetected features. Once the procedure goes through the fourth to sixth images, the focused features are near the face area. In addition, body images were obtained from the batch-normalized layer. Contrary to the results of the face images, the major characteristics were observed near the area of the body images since the trained part of the ResNet-50 model regards data that pertain to the person's clothes and body as important features.

Figure 21. Results on class activation feature map on ChokePoint dataset [34]. (a,b) are activation map results from face images. From the left image to the right image: original image, low-illumination image, enhancement image from modified EnlightenGAN, result from 7th ReLU layer, result from 12th ReLU layer, and result from 13th ReLU layer of VGG face-net 16. (c,d) are activation map results from body images. From the left image to the right image: original image, low-illumination image, enhancement image from modified EnlightenGAN, result from 3rd batch-normalized layer, result from conv5 2nd block, and result from conv5 3rd block of ResNet-50.

4.5. Testing of Proposed Method with Low-Illuminated Images Captured in Real Environments

In our experiments, we compared the performance by the proposed modified EnlightenGAN with that by the original EnlightenGAN in terms of the quality of the restored

images, as shown in Table 7. The quality of the restored images was calculated based on the similarity between the original and restored images. Therefore, the ground truth original images were necessary for our experiments, and we obtained the simulated low-light images from the original ones. Nevertheless, we performed the first additional experiments with low-illuminated images which were not simulated but captured in real low-light environments. This database was captured from 20 people [52], and most of the participants were the same as those in our first database of DFB-DB3 in order for us to consider only the factor of the simulated and real low-illumination images, while other factors were kept similar for these two databases for fair comparisons. Figure 22 shows examples of original images captured in real low-light environments. The total number of face images is 10,095 and that of body images is also 10,095, which were captured from 20 people [52]. Images were acquired by a visible-light camera in a real surveillance camera environment. The height of the camera was about 2.3 m from the ground, and the distance between the camera and a person was about 20~22 m. Images were acquired in a real night environment of about 10~20 lux. (at 9~10 p.m.) [52]. In our experiments, we performed two-fold cross-validation. In the first fold, 4286 images of the face or body from 10 people were used for training, and the remaining 5809 images of the face or body from the other 10 people were used for testing. In the second fold, 5809 images of the face or body from 10 people were adopted for training, and the remaining 4286 images of the face or body from the other 10 people were used for testing. From these, we determined the average value of the two testing accuracies as the final accuracy. As shown in Table 18 and Figure 23, we compared the accuracies of face and body recognition by the proposed method with those by the state-of-the-art methods. As shown in this table and figure, our method outperforms the state-of-the-art methods. In addition, the accuracies by our method are similar to those using the simulated images of Table 12, Figures 10b and 11b, which confirms that our method can be applied to low-illuminated images captured in a real environment.

Figure 22. Examples of original images captured in real low-light environments.

Table 18. Comparison of EERs using the proposed method and state-of-the-art body and face recognition methods (unit: %).

Method	EER
ResNet IDE + LIME [47]	21.17
ELF [48]	32.65
Proposed method	7.39

Figure 23. Comparisons of (a) ROC and (b) CMC curves obtained by proposed method and state-of-the-art-methods.

In addition, we performed the second additional experiments with an open database from Fudan University [53]. This open database includes the images captured at Fudan University in a real low-light environment, as shown in Figure 24. The total number of images is 275, and the images were captured from six people with four~six people per image. As with the experiments with the first database of real low-light environments [52], we performed two-fold cross-validation. In the first fold, 708 images of the face or body from three people were used for training, and the remaining 711 images of the face or body from the other three people were used for testing. In the second fold, 711 images of the face or body from three people were adopted for training, and the remaining 708 images of the face or body from the other three people were used for testing. As the number of training data is relatively smaller than that of the first database [52], we performed the fine-tuning of our model trained with the first database for training. From these, we determined the average value of the two testing accuracies as the final accuracy, and the final accuracies shown in Table 19 and Figure 25 confirm that our method outperforms the state-of-the-art methods. In addition, the accuracies by our method are similar to those using the simulated images of Table 12, Figures 10b and 11b, which confirms that our method can be applied to low-illuminated images captured in a real environment.

Figure 24. Example of an original image from an open database captured in a real low-light environment.

(a)

(b)

Figure 25. Comparisons of (a) ROC and (b) CMC curves obtained by the proposed method and state-of-the-art-methods with an open database.

Table 19. Comparison of EERs using the proposed method and state-of-the-art face and body recognition methods with an open database (unit: %).

Method	EER
ResNet IDE + LIME [47]	22.54
ELF [48]	31.78
Proposed method	7.32

4.6. Comparisons of Desktop Computer and Jetson TX2 Processing Time

For the following experiment, the proposed method for the computational speed was compared by applying a Jetson TX2 board [54], as shown in Figure 26, and the desktop computer discussed in Section 4.1. The Jetson TX2 board is an embedded system installed with an NVIDIA Pascal™ graphics processing unit (GPU) structure with 8-GB 128-bit LPDDR4 memory, a dual-core NVIDIA Denver 2 64-Bit CPU, and 256 NVIDIA CUDA cores. The power consumption is lower than 7.5 W. The proposed method was installed with TensorFlow [55] and Keras [56] in Ubuntu 16.04 OS. The installed library and framework versions include TensorFlow 1.12 and Python 3.5; the NVIDIA CUDA® deep neural network library (CUDNN) [57] and NVIDIA CUDA® toolkit [58] are each 7.3 and 9.0.

Figure 26. Jetson TX2 embedded system.

The processing time for the modified EnlightenGAN, and the processing speed on a desktop computer were observed to be faster than those in the Jetson TX2, as shown Table 20. Table 21 presents the processing speed of the CNNs for recognition. Based on Tables 20 and 21, we can confirm that our method can be applied to both desktop computers and embedded systems having low processing powers and resources.

Table 20. Comparison of desktop computer and Jetson TX2 processing times by EnlightenGAN (unit: ms).

GAN Model	Jetson TX2	Desktop Computer
Modified EnlightenGAN	288.1	18.8

Table 21. Comparison of desktop computer and Jetson TX2 processing times by ResNet-50 and VGG face net-16 (unit: ms).

CNN Model	Desktop Computer	Jetson TX2
VGG face net-16	24.6	91.7
ResNet-50	18.93	40.9
Total	43.53	132.6

5. Conclusions

This study proposed the modified EnlightenGAN to solve the problem of low illumination, which occurs during long-distance human recognition in a dark indoor environment, in which score-level fusion was applied to the CNN matching scores of face and body images for recognition. In previous studies, recognition was performed using face images captured in environments where low illumination is not severe, whereas this study focused on improving the face and body recognition of images obtained from very low illumination environments. To address this issue, this study used the modified EnlightenGAN to independently perform low-illumination enhancement for face images and body images. Although independent operation would increase the processing time of the GAN by a factor of two, it can prevent the loss of important facial and body features due to low illumination. Based on the experiment conducted with two databases, the proposed method exhibited a better recognition performance than the state-of-the-art methods and was feasible in a desktop computer as well as a Jetson TX2 embedded system. Moreover, significant features for recognition were detected adequately in CNN models that use the images enhanced by the modified EnlightenGAN through a class activation map. For a fair comparison, the use of DFB-DB3, the modified EnlightenGAN, and deep CNNs for human recognition is possible by other researchers through the website of [59].

In future work, the applicability of the modified EnlightenGAN-based recognition method proposed in this study to various biometric recognition systems, such as iris and gait recognition, will be evaluated. A GAN model that is lighter and that is processed faster than the proposed GAN model will be developed to increase the processing speed in an embedded system environment.

Author Contributions: Methodology, J.H.K.; conceptualization, S.W.C.; validation, N.R.B.; supervision, K.R.P.; writing—original draft, J.H.K.; writing—editing and review, K.R.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (MSIT) through the Basic Science Research Program (NRF-2021R1F1A1045587), in part by the NRF funded by the MSIT through the Basic Science Research Program (NRF-2019R1A2C1083813), and in part by the MSIT, Korea, under the ITRC (Information Technology Research Center) support program (IITP-2021-2020-0-01789) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: All subjects gave their informed consent for inclusion before they participated in the study.

Data Availability Statement: Not applicable.

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

References

1. Grgic, M.; Delac, K.; Grgic, S. SCface—surveillance cameras face database. *Multimed. Tools Appl.* **2011**, *51*, 863–879. [[CrossRef](#)]
2. Banerjee, S.; Das, S. Domain adaptation with soft-margin multiple feature-kernel learning beats deep learning for surveillance face recognition. *arXiv* **2016**, arXiv:1610.01374v2.
3. Varior, R.R.; Haloi, M.; Wang, G. Gated siamese convolutional neural network architecture for human re-identification. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 791–808.
4. Shi, H.; Yang, Y.; Zhu, X.; Liao, S.; Lei, Z.; Zheng, W.; Li, S.Z. Embedding deep metric for individual re-identification: A study against large variations. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 732–748.
5. Han, J.; Bhanu, B. Statistical feature fusion for gait-based human recognition. In Proceedings of the IEEE Conference and Computer Vision and Pattern Recognition, Washington, DC, USA, 27 June–2 July 2004; pp. II-842–II-847.
6. Liu, Z.; Sarkar, S. Outdoor recognition at a distance by fusing gait and face. *Image Vision Comput.* **2007**, *25*, 817–832. [[CrossRef](#)]
7. Koo, J.H.; Cho, S.W.; Baek, N.R.; Kim, M.C.; Park, K.R. CNN-based multimodal human recognition in surveillance environments. *Sensors* **2018**, *18*, 3040. [[CrossRef](#)] [[PubMed](#)]

8. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–14.
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
10. Kamenetsky, D.; Yiu, S.Y.; Hole, M. Image enhancement for face recognition in adverse environments. In Proceedings of the Digital Image Computing: Techniques and Applications, Canberra, Australia, 10–13 December 2018; pp. 1–6.
11. Huang, Y.H.; Chen, H.H. Face recognition under low illumination via deep feature reconstruction network. In Proceedings of the IEEE International Conference on Image Processing, Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 2161–2165.
12. Poon, B.; Amin, M.A.; Yan, H. PCA based human face recognition with improved methods for distorted images due to illumination and color background. *IAENG Intern. J. Comput. Sci.* **2016**, *43*, 277–283.
13. Zhang, T.; Tang, Y.Y.; Fang, B.; Shang, Z.; Liu, X. Face recognition under varying illumination using gradientfaces. *IEEE Trans. Image Proc.* **2009**, *18*, 2599–2606. [[CrossRef](#)] [[PubMed](#)]
14. Zhao, M.; Wang, L. Face recognition based on a novel illumination normalization method. In Proceedings of the 5th International Congress on Image and Signal Processing, Chongqing, China, 16–18 October 2012; pp. 434–438.
15. Vu, N.S.; Caplier, A. Illumination-robust face recognition using retina modeling. In Proceedings of the 16th IEEE International Conference on Image Processing, Cairo, Egypt, 7–10 November 2009; pp. 3289–3292.
16. Kang, Y.; Pan, W. A novel approach of low-light image denoising for face recognition. *Adv. Mech. Eng.* **2014**, *6*, 1–13. [[CrossRef](#)]
17. Ren, D.; Ma, H.; Sun, L.; Yan, T. A novel approach of low-light image used for face recognition. In Proceedings of the 4th International Conference on Computer Science and Network Technology, Harbin, China, 19–20 December 2015; pp. 790–793.
18. Viola, P.; Jones, M.J. Robust Real-Time Face Detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [[CrossRef](#)]
19. Kazemi, V.; Sullivan, J. OneMillisecond Face Alignment with an Ensemble of Regression Trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1867–1874.
20. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *arXiv* **2014**, arXiv:1406.2661.
21. Kupyn, O.; Budzan, V.; Mykhailych, M.; Mishkin, D.; Matas, J. DeblurGAN: Blind motion deblurring using conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8183–8192.
22. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976.
23. Zhu, J.-Y.; Park, T.S.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2242–2251.
24. Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; Wang, Z. EnlightenGAN: Deep light enhancement without paired supervision. *IEEE Trans. Image Process.* **2021**, *30*, 2340–2349. [[CrossRef](#)] [[PubMed](#)]
25. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
26. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 448–456.
27. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.K.; Wang, Z.; Smolley, S.P. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2813–2821.
28. Wolf, L.; Hassner, T.; Maoz, I. Face recognition in unconstrained videos with matched background similarity. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 529–534.
29. Huang, G.B.; Ramesh, M.; Berg, T.; Learned-miller, E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In Proceedings of the Workshop on Faces in ‘Real-Life’ Images: Detection, Alignment, and Recognition, Marseille, France, 17 October 2008; pp. 1–11.
30. Mateo, J.R.S.C. Weighted Sum Method and Weighted Product Method. In *Multi Criteria Analysis in the Renewable Energy Industry*; Green Energy and Technology; Springer: London, UK, 2012. [[CrossRef](#)]
31. Vapnik, V. *Statistical Learning Theory*; Wiley: Hoboken, NJ, USA, 1998.
32. Logitech BCC950 Camera. Available online: <https://www.logitech.com/en-roeu/product/conferencecam-bcc950?crd=1689> (accessed on 1 March 2021).
33. Logitech C920 Camera. Available online: <https://www.logitech.com/en-us/product/hd-pro-webcam-c920?crd=34> (accessed on 1 March 2021).
34. ChokePoint Dataset. Available online: <http://arma.sourceforge.net/chokepoint/> (accessed on 26 February 2021).
35. CUDA. Available online: <https://developer.nvidia.com/cuda-10.0-download-archive> (accessed on 18 April 2021).
36. NVIDIA GeForce GTX 1070 Card. Available online: <https://www.nvidia.com/en-in/geforce/products/10series/geforce-gtx-1070/> (accessed on 12 May 2021).
37. Pytorch. Available online: <https://pytorch.org/get-started/previous-versions> (accessed on 18 April 2021).

38. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 675–678.
39. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
40. Oxford Face Database. Available online: https://www.robots.ox.ac.uk/~lvgg/data/vgg_face/ (accessed on 18 April 2021).
41. Stathaki, T. *Image Fusion: Algorithms and Applications*; Academic Press: Cambridge, MA, USA, 2008.
42. Salomon, D. *Data Compression: The Complete Reference*, 4th ed.; Springer: New York, NY, USA, 2006.
43. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]
44. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep face recognition. In Proceedings of the British Machine Vision Conference, Swansea, UK, 7–10 September 2015; pp. 1–12.
45. Gruber, I.; Hlaváč, M.; Železný, M.; Karpov, A. Facing face recognition with ResNet: Round one. In Proceedings of the International Conference on Interaction Collaborative Robotics, Hatfield, UK, 12–16 September 2017; pp. 67–74.
46. Martínez-Díaz, Y.; Méndez-Vázquez, H.; López-Avila, L.; Chang, L.; Enrique Sucar, L.; Tistarelli, M. Toward more realistic face recognition evaluation protocols for the youtube faces database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 526–534.
47. Guo, X.; Li, Y.; Ling, H. LIME: Low-Light Image Enhancement via Illumination Map Estimation. *IEEE Trans. Image Process.* **2017**, *26*, 982–993. [[CrossRef](#)] [[PubMed](#)]
48. Gray, D.; Tao, H. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In Proceedings of the 10th European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 262–275.
49. Livingston Edward, H. Who was student and why do we care so much about his *t*-test? *J. Surg. Res.* **2004**, *118*, 58–65. [[CrossRef](#)] [[PubMed](#)]
50. Cohen, J. A power primer. *Psychol. Bull.* **1992**, *112*, 155–159. [[CrossRef](#)] [[PubMed](#)]
51. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
52. Cho, S.W.; Baek, N.R.; Kim, M.C.; Koo, J.H.; Kim, J.H.; Park, K.R. Face Detection in Nighttime Images Using Visible-Light Camera Sensors with Two-Step Faster Region-Based Convolutional Neural Network. *Sensors* **2018**, *18*, 2995. [[CrossRef](#)] [[PubMed](#)]
53. Open Database of Fudan University. Available online: https://cv.fudan.edu.cn/_upload/tpl/06/f4/1780/template1780/humandetection.htm (accessed on 26 May 2021).
54. Jetson TX2 Module. Available online: <https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-tx2/> (accessed on 12 December 2020).
55. Tensorflow: The Python Deep Learning Library. Available online: <https://www.tensorflow.org/> (accessed on 12 December 2020).
56. Keras: The Python Deep Learning Library. Available online: <https://keras.io/> (accessed on 12 December 2020).
57. CUDNN. Available online: <https://developer.nvidia.com/cudnn> (accessed on 11 January 2021).
58. CUDA. Available online: <https://developer.nvidia.com/cuda-90-download-archive> (accessed on 11 January 2021).
59. Dongguk Face and Body Database Version 3 (DFB-DB3), Modified EnlightenGAN, and CNN Models for Face & Body Recognition. Available online: <http://dm.dgu.edu/link.html> (accessed on 11 March 2021).