mathematics

MDPI

Article

Pedestrian Gender Recognition by Style Transfer of Visible-Light Image to Infrared-Light Image Based on an Attention-Guided Generative Adversarial Network

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

Division of Electronics and Electrical Engineering, Dongguk University, 30 Pildong-ro 1-gil, Jung-gu, Seoul 04620, Korea; naris27@dgu.ac.kr (N.R.B.); jsu319@dongguk.edu (S.W.C.); koo6190@dongguk.edu (J.H.K.)
* Correspondence: parkgr@dgu.edu; Tel.: +82-10-3111-7022; Fax: +82-2-2277-8735

Abstract: Gender recognition of pedestrians in uncontrolled outdoor environments, such as intelligent surveillance scenarios, involves various problems in terms of performance degradation. Most previous studies on gender recognition examined recognition methods involving faces, full body images, or gaits. However, the recognition performance is degraded in uncontrolled outdoor environments due to various factors, including motion and optical blur, low image resolution, occlusion, pose variation, and changes in lighting. In previous studies, a visible-light image in which image restoration was performed and infrared-light (IR) image, which is robust to the type of clothes, accessories, and lighting changes, were combined to improve recognition performance. However, a near-IR (NIR) image requires a separate NIR camera and NIR illuminator, because of which challenges are faced in providing uniform illumination to the object depending on the distance to the object. A thermal camera, which is also called far-IR (FIR), is not widely used in a surveillance camera environment because of expensive equipment. Therefore, this study proposes an attention-guided GAN for synthesizing infrared image (SI-AGAN) for style transfer of visible-light image to IR image. Gender recognition performance was improved by using only a visible-light camera without an additional IR camera by combining the synthesized IR image obtained by the proposed method with the visible-light image. In the experiments conducted using open databases—RegDB database and SYSU-MM01 database—the equal error rate (EER) of gender recognition of the proposed method in each database was 9.05 and 12.95%, which is higher than that of state-of-the-art methods.

Keywords: gender recognition; visible-light and IR cameras; style transfer of visible-light image to IR image; SI-AGAN

Citation: Baek, N.R.; Cho, S.W.; Koo, J.H.; Park, K.R. Pedestrian Gender Recognition by Style Transfer of Visible-Light Image to Infrared-Light Image Based on an Attention-Guided Generative Adversarial Network. *Mathematics* **2021**, *9*, 2535. https://doi.org/10.3390/math9202535

Academic Editors: George E. Tsekouras, Christos Kalloniatis and Dimitrios Makris

Received: 13 September 2021
Accepted: 28 September 2021
Published: 9 October 2021

1. Introduction

Pedestrian gender recognition in uncontrolled environments has been considered across an array of fields, such as computer vision, marketing, security surveillance, forensic affairs, and human–robot interactions. Conventional gender recognition software recognizes genders based on high-resolution facial images captured in a controlled environment [1] or based on continuously imaged gaits [2,3]. However, images acquired from uncontrolled environments significantly reduce gender recognition performance due to low image resolution, occlusion, images of backside appearance, lighting changes, and optical and motion blur. Gender recognition has been performed using full body images of a person in uncontrolled environments [4]. Pedestrian gender recognition using full body images has limited recognition performance due to the following challenging points. First, gender recognition using a full body image is sensitive to a person's hair style or clothes [4]. For images of a person captured from behind, gender recognition is performed based on a person's hair style or clothes. However, it is difficult to distinguish between two genders if the person is wearing unisex clothes. In particular, it is even more challenging to discern the gender from images taken in winter due to thick padded coats. Male and

female subjects also may have similar hair styles, and it is difficult to discern their genders if the subjects are wearing hats or caps. Second, the images used for pedestrian gender recognition often have motion blur, optical blur, and noise as they are captured from a distance. Therefore, most pedestrian images are low-resolution images, thus degrading gender recognition performance. Third, images also have occlusion, pose changes, and illumination variation as they are captured in uncontrolled environments. The performance of pedestrian gender recognition is typically limited due to these challenges.

Most previous studies on pedestrian gender recognition performed gender recognition using only visible-light images [4–10]. However, gender recognition based on visible-light images has reduced recognition performance because features are difficult to train when training a recognizer such as convolutional neural network (CNN) due to excessive information such as background, accessories, clothes, and hair styles. To overcome such drawbacks, infrared (IR) images were combined with visible-light images to enhance the recognition performance in a previous study [11]. For using the gender recognition method of the study [11], both visible-light images and IR images are required during testing. However, near-IR (NIR) images require a separate NIR camera and NIR lighting, which faces issues while providing uniform illumination to the object depending on the distance from the object. A thermal camera, also called far-IR (FIR) camera, is not widely used in a surveillance environment because of expensive equipment. Therefore, this study proposes an attention-guided generative adversarial network (GAN) for synthesizing infrared image (hereinafter called SI-AGAN), which performs style transfer from a visible-light image to a synthesized IR image (syn-IR image) through a GAN. The gender recognition performance was enhanced by combining syn-IR images generated through SI-AGAN and improved visible-light images obtained by sequentially running two CNN models. This study has the following four contributions:

- For improving gender recognition performance, we propose SI-AGAN, which transfers the style of the visible-light image to resemble that of an IR image. Existing multimodal camera-based methods required both a visible-light image and an IR image during training and testing. In this study, however, an IR image is not required during testing as the IR image generated by SI-AGAN is used.
- We reduced the computational cost of the SI-AGAN by revising convolutional layers of the attention module, attention-guided generator, and attention-guided discriminator of the original attention-guided generative adversarial network (AGGAN) to a depthwise separable convolution layer.
- Furthermore, the quality of generated images and the gender recognition performance were improved by applying a perceptual loss in SI-AGAN. Moreover, the matching score obtained through the residual network (ResNet)-101, trained with a visible-light image and the syn-IR image generated by SI-AGAN, was applied with score-level fusion based on a support vector machine (SVM) to improve gender recognition performance.
- Our trained SI-AGAN models and the generated syn-IR dataset are disclosed through [12] for a fair performance evaluation by other researchers.

The remaining parts of this paper are organized as follows. Section 2 highlights the previous studies on pedestrian gender recognition. Section 3 explains our proposed method, whereas Section 4 describes the experiment results and analysis. Finally, Section 5 concludes our study.

2. Related Work

In previous studies on gender recognition, face-based gender recognition was mostly performed using clear facial images captured from a close distance [13]. In uncontrolled environments, such as an intelligent surveillance system, however, detecting facial images is challenging because the images are captured from a distance or detecting the face is difficult in occluded images or in images taken from the side or behind. Due to such issues, human body images have been used in previous studies on pedestrian gender recognition.

Extensive research has been conducted on extracting features for recognizing genders in pedestrian images. The extracted features can be divided into handcrafted feature-based or deep feature-based approaches according to the extraction method.

2.1. Handcrafted Feature-Based Methods

Pedestrian gender recognition commonly uses images captured from a distance in uncontrolled environments. Hence, the images are noisy and blurry. Previously, several studies have been conducted on gender recognition in which color information of clothes was used based on handcrafted features. Additional research was conducted on using different handcrafted features for each view by distinguishing the front view or back view. In the first study on pedestrian gender recognition [4], a pedestrian image was segmented into patch images from which the histogram of oriented gradients (HOGs) feature vector [14] was extracted to perform classification through adaptive boosting (Adaboost) [15] and random forest [16] methods. An edge map was used instead of a raw pixel, considering the changes in the color of clothes in the study [4]. However, color information is particularly important in studies on pedestrian gender recognition. Thus, in [5], shape features obtained by pixelHOG (PiHOG) and color features obtained through local HSV color histogram (LSHV) were combined, and gender recognition was performed using a linear-kernel SVM [17]. However, color features obtained through LSHV have insufficient color representation, and the experiment was only conducted for the frontal view, thus making it difficult to apply to back or side view images. In [6], gender recognition was performed by extracting features using a part-based method based on poselets to make the model robust to camera view. In [7], biologically inspired features (BIF) extracted through a Gabor filter were combined with handcrafted features through principal component analysis (PCA) [18], orthogonal locality preserving projections [19], locality sensitive discriminant analysis (LSDA) [20], and marginal fisher analysis [21] to classify the frontal and back view. However, performance evaluation was not conducted for the side view, because PCA and LSDA need to be performed separately for each view. In [22], gender recognition was performed using pedestrian images in which thermal images were used in addition to visible-light images. HOG features were extracted from visible-light images and thermal images and then combined to perform gender recognition. However, gender recognition performance may be degraded due to the effects of the background region. In [23], a weight HOG was proposed in which a greater weight was given to the bright region when extracting HOG features in a visible-light image based on the fact that the human region, which is the object of thermal light, is brighter than the background region.

2.2. Deep Feature-Based Methods

In handcrafted feature-based pedestrian gender recognition, features are extracted using pre-designed HOG, PiHOG, and BIF. Then, gender recognition is performed with a separate classifier, SVM. It is difficult to respond flexibly to various types of data or circumstances as pre-designed and fixed features are used. Therefore, research is actively being conducted on pedestrian gender recognition through a deep feature-based method, where features are self-extracted during training.

2.2.1. CNN-Based Methods

Starting with [24], extensive research has been conducted on various object recognition techniques using a CNN in which features are automatically trained in training data and no separate classifier is required, unlike handcrafted feature-based methods, and notable performances have been observed. Subsequently, CNNs provide increasingly superior performance in pedestrian gender recognition. In [8], a CNN was used in pedestrian gender recognition. More outstanding or similar performance as the conventional handcrafted feature-based methods was observed with a simple architecture. In [9], gender recognition was performed using Mini-CNN and AlexNet [24]. When deep feature and HOG feature

were compared through a homogeneous dataset and heterogeneous dataset, a deep feature was proven to show better performance, especially for the heterogeneous datasets. In [10], a global CNN was trained using whole-body-part images, while the remaining three parts of the body were used to train a local CNN each; ultimately, four CNNs were combined. CaffeNet [25] and visual geometry group (VGG) Net-19 [26] were used as the CNN. In [27], the authors claimed that the background region of an image is the cause of performance degradation in pedestrian gender recognition. Hence, a stacked sparse auto encoder (SSAE) was proposed for removing the background region as a preprocessing step for a pedestrian image. Only a CNN has been used thus far in performing gender recognition. Deep-learned features were combined with the weighted HOG handcrafted features in [28]. After generating a fusion layer by applying fusion to two features, the Softmax classifier was used to perform gender recognition. Deep-learned and handcrafted features were also combined in [29]. Better recognition performance was shown through a joint feature acquired by combining obtained deep-learned features through VGG Net-19 and a deep ResNet [30], and local maximal occurrence (LOMO) [31] features and HOG features, which are handcrafted features.

Previous studies that have employed a CNN only used visible-light images. However, Ref. [32] stated that the visible light image is sensitive in various environments for pedestrian recognition, so it will be helpful for performance as a multimodal camera-based method. In [33], features were extracted by training AlexNet with visible-light image and thermal image separately, and gender recognition was performed through SVM. In [11], the gender recognition performance was enhanced by combining IR image with visible-light image improved through CNN-based two-step reconstruction.

2.2.2. CNN and GAN-Based Methods

Starting with [34], a GAN has been widely used across various fields, such as style transfer, augmentation, super resolution, and image completion. A GAN consists of a generator and a discriminator. A generator generates fake images that appear real, whereas the discriminator discriminates real images from fake images. Several studies are being conducted in which a GAN is used for improving performance in pedestrian gender recognition. The authors of [35] proposed a key pedestrian transfer generative adversarial network (KPT-GAN). This network is designed to be robust to background changes by applying scene transfer to the background. Moreover, gender recognition is performed through CNN-based viewpoint adaptation feature learning. In most previous studies on pedestrian gender recognition, the background region is removed in the pedestrian image using various techniques or the background is changed through scene transfer using a GAN to improve pedestrian gender recognition performance. In addition, pedestrian gender recognition entails degraded recognition performance due to motion blur, optical blur, and sensor noise as the images are captured from a distance. Considering these drawbacks, this study proposes a gender recognition method in which a visible-light image for which blur and noise are improved through two-step reconstruction is fused with a syn-IR image, which is generated through SI-AGAN to be similar to an IR image that is less affected by background, shadow, lighting changes, clothing type, and accessories. Table 1 presents a comparison of the advantages and disadvantages between the proposed method and previous methods on pedestrian gender recognition.

Table 1. Comparison of the previous and the proposed methods for pedestrian gender recognition.

Category			Advantage	Disadvantage
Handcrafted feature	HOG [4]		Effectively extracts HOG features by segmenting a pedestrian image into patches	Limited recognition performance from not using color information of a person
	PiHOG + LSHV [5]		The color feature is used to effectively extract gender features	Gender recognition cannot be performed using back or side view images
	Poselets [6]		Robust to occlusion images by using a part-based method	Requires heavily annotated information
	BIFs [7]		View classification is performed through BIFs with manifold learning and gender recognition is performed for each view	No performance available for side view
	Multimodal Camera-based HOG [22,23]		Visible-light image and thermal image are combined after applying HOG or weighted HOG	Less accurate recognition because of low-resolution images captured from a far distance
Deep feature	CNN	CNN [8]	Similar or better performance than a handcrafted feature with simple architecture	Cannot exhibit high performance improvement due to a simple architecture
		Mini-CNN [9]	Excellent performance in heterogeneous datasets compared to handcrafted features	Similar performance as a handcrafted feature in a homogeneous dataset
		Global CNN + local CNN [10]	More discriminative features can be obtained by separately training the global region and local region	Complicated due to the use of four CNNs and can only be applied to a whole image
		SSAE [27]	Performance is improved by using a pedestrian image in which the background is removed through preprocessing	Limited recognition performance as color information, which is an important factor
		Handcrafted feature + CNN [28,29]	The discriminative features can be obtained by combining low-level and high-level features	Limited recognition performance due to various factor as only a visible-light image is used
		Multimodal Camera-based CNN [11,33]	Less affected by background, lighting changes, clothing type, and accessories because visible-light image and thermal image are combined	Time-consuming since a CNN is applied to the visible-light image and IR image separately
	CNN + GAN	KPT-GAN [35]	Robust to scene variation as KTP-GAN that performs scene transfer for the background is used	Numerous artifacts exist, and the scene transferred background region is noisier than the actual background
		SI-AGAN (Proposed method)	Recognition performance is improved by combining reconstructed visible-light images with the syn-IR image generated by the proposed SI-AGAN to be similar as an IR image	Two-step image reconstruction and GAN are time-consuming

3. Proposed Method

3.1. Overview of the Proposed Method

Figure 1 shows the overall flowchart of the proposed methods and Algorithm 1 shows the pseudo cod of the proposed methods. In step (1), a visible-light pedestrian image is acquired in uncontrolled environments. The acquired visible-light pedestrian image is blurry and noisy because the objects were moving pedestrians in uncontrolled environments. In step (2), two-step image reconstruction is performed to improve image

quality. In step (3), the score of gender recognition is obtained using the improved image as the input of a deep ResNet. When only the visible-light image is used for pedestrian gender recognition, the performance can be reduced due to the background of a pedestrian image or hair style, accessories, and clothes of a pedestrian [11,22,23,33]. Therefore, a syn-IR image, which is similar to an IR image where a human region is distinctive, is generated through SI-AGAN to perform pedestrian gender recognition. Accordingly, in step (4), the image is converted to a grayscale image to reduce the influence of color information in the visible-light image, and then the converted image is used as the input of SI-AGAN to obtain a syn-IR image. In step (5), the score of gender recognition is obtained by using the syn-IR image as the input of a deep ResNet. In step (6), score fusion is applied through SVM to the scores obtained from each visible-light image, and the gender is finally determined in step (7). Two-step image reconstruction is further explained in Section 3.2, the proposed SI-AGAN for generating vis-image in Section 3.3, and SVM-based score-level fusion in Section 3.4.

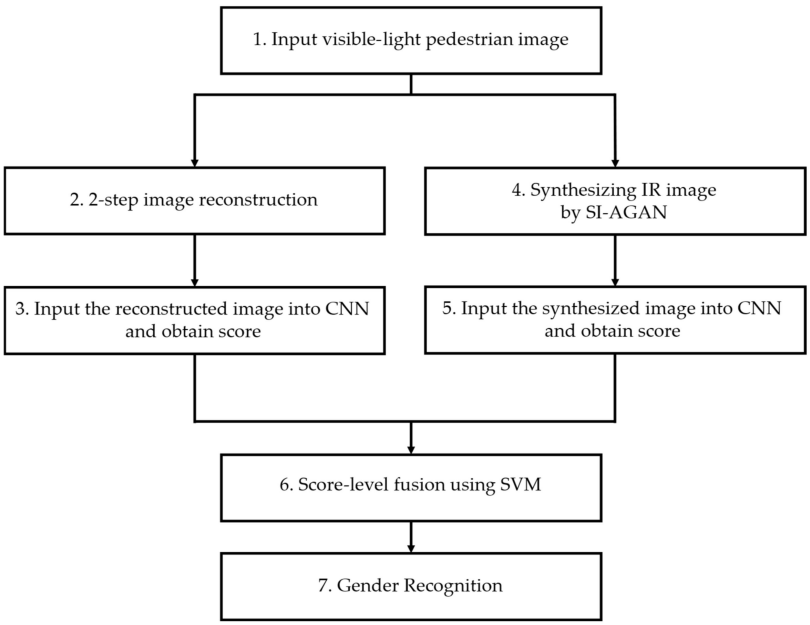


Figure 1. Flowchart of the proposed method.

Algorithm 1 The proposed method detailed by using pseudo code

Input visible-light image: $X = \{X_1, \ldots X_m\}$
Input reconstructed image: $R = \{R_1, \ldots R_m\}$
Input synthesized image: $S = \{S_1, \ldots S_m\}$
Output score obtained from reconstructed image: $O = \{O_1, \ldots, O_m\}$
Output score obtained from synthesized image: $N = \{N_1, \ldots, N_m\}$
Final output score: Z
2-step image reconstruction model = *r_model*
SI-AGAN model = s_*model*
Gender recognition CNN model = *g_model*
SVM classifier = *svm*

Algorithm procedure

$X = \{X_{t-0},\ X_{t-1}, \ldots\ X_t\}$
$R = \{R_1, \ldots R_m\} = r\_model(X)$
$S = \{S_1, \ldots S_m\} =$ s_$model(X)$

$O = \{O_1, \ldots, O_m\} = g\_model(R)$
$N = \{N_1, \ldots, N_m\} = g\_model(S)$

$O\prime = concatenate(O_1, \ldots, O_m,\ axis = vertical)$
$N\prime = concatenate(N_1, \ldots, N_m,\ axis = vertical)$

$Z = svm(O',\ N')$

3.2. 2-Step Image Reconstruction

Pedestrian images captured in uncontrolled environments have degraded recognition performance due to optical blur, motion blur, noise, and low resolution. To solve such problems, enhanced performance was achieved by improving the visible-light image through CNN-based two-step image reconstruction in [11]; the quality of the visible-light image was also improved in this study through CNN-based two-step image reconstruction. The two-step image reconstruction process is as follows. In the first step, denoising is performed using an image restoration CNN (IRCNN) [36]. An IRCNN is a residual learning-based method in which the noise information of an input image is learned and subtracted. The architecture of an IRCNN is explained in Table 2. Then, in the second step, the image quality is enhanced through super-resolution using very deep convolutional networks (VDSR) [37]. The VDSR learns the shape information of an input image and adds it to the input image. The architecture of VDSR is explained in Table 3. Using two types of CNNs, the improved image quality is obtained by adding the shape information and removing noise in images captured under uncontrolled environments. A further explanation is provided below.

Table 2. Architecture of IRCNN. (D Conv in n-D Conv indicates a dilated convolution layer. Here, n is the dilation rate, which is the same as that applied for standard convolution when n = 1. ReLU refers to a rectified linear unit, and Bnorm refers to batch normalization. An IRCNN uses an original image with unfixed width and height; thus, W and H are denoted).

Layer Type	Number of Filters	Size of Feature Map (Width × Height × Channel)	Size of Kernel (Width × Height)	Number of Stride	Number of Padding
Input layer [image]		W × H × 3			
1-D Conv 1 (ReLU)	64	W × H × 64	3 × 3	1 × 1	1 × 1
2-D Conv 2 (Bnorm + ReLU)	64	W × H × 64	3 × 3	1 × 1	2 × 2
3-D Conv 3 (Bnorm + ReLU)	64	W × H × 64	3 × 3	1 × 1	3 × 3
4-D Conv 4 (Bnorm + ReLU)	64	W × H × 64	3 × 3	1 × 1	4 × 4
3-D Conv 5 (Bnorm + ReLU)	64	W × H × 64	3 × 3	1 × 1	3 × 3
2-D Conv 6 (Bnorm + ReLU)	64	W × H × 64	3 × 3	1 × 1	2 × 2
1-D Conv 7	64	W × H × 64	3 × 3	1 × 1	1 × 1

Table 3. Architecture of VDSR. (Conv refers to a convolution layer. Here, N* indicates a number from 1 to 19. ReLU refers to a rectified linear unit. VDSR uses an original image with unfixed width and height; thus, W and H are denoted).

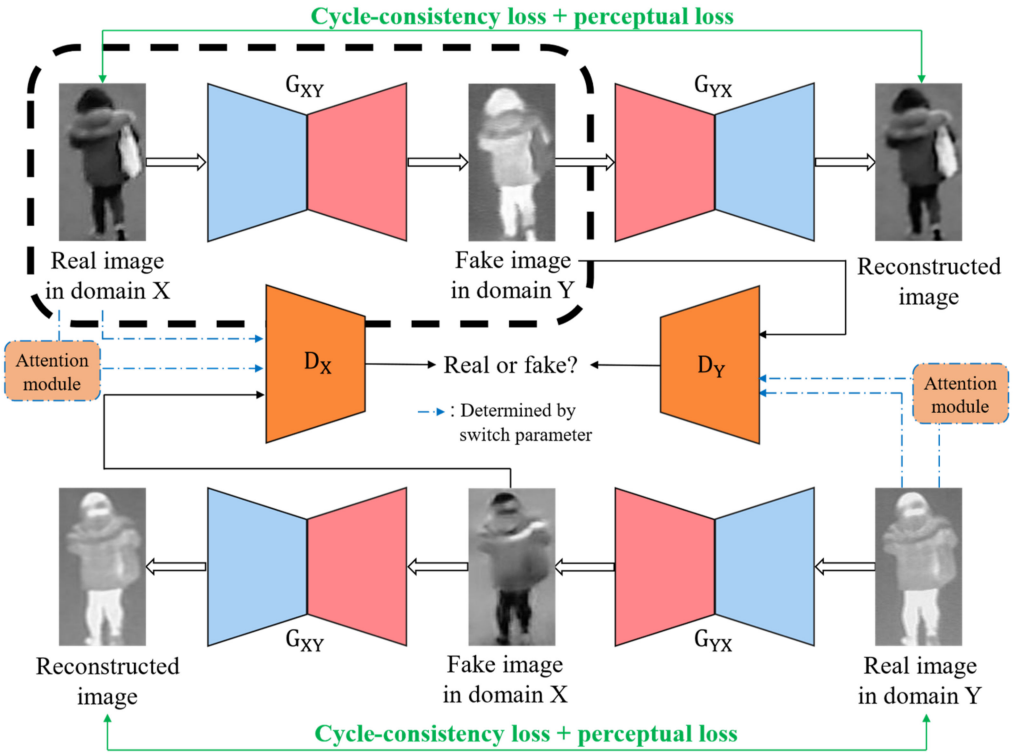
Layer Type	Number of Filters	Size of Feature Map (Width × Height × Channel)	Size of Kernel (Width × Height)	Number of Stride	Number of Padding
Input layer [image]		W × H × 3			
Conv N* (ReLU)	64	W × H × 64	3 × 3	1 × 1	1 × 1
Conv 20	64	W × H × 64	3 × 3	1 × 1	1 × 1

When the two CNNs were applied in this study, we used pre-trained models rather than separately training them with the training dataset. Because the dataset used in this study was captured in uncontrolled environments, there is no pair of low-resolution, noisy images and high-resolution, denoised images.

3.3. SI-AGAN

The shape information of a person is as important as a person's hair style or clothing in terms of improving the performance of pedestrian gender recognition. However, pedestrian gender recognition using only a visible-light image shows poor performance because feature extraction from an image focus on the background, a person's hair style, accessories, and clothes. To overcome this drawback, this study proposes SI-AGAN, which generates a syn-IR image in which pedestrian region information is considered important in an IR image.

SI-AGAN is largely divided into a generator and a discriminator, as shown in Figure 2, and an attention module is added to the generator. The generator of SI-AGAN is further explained in Section 3.3.1, in addition to the attention module. The discriminator of SI-AGAN is further explained in Section 3.3.2, while the loss of SI-AGAN is further explained in Section 3.3.3.



(**a**)

Figure 2. *Cont.*

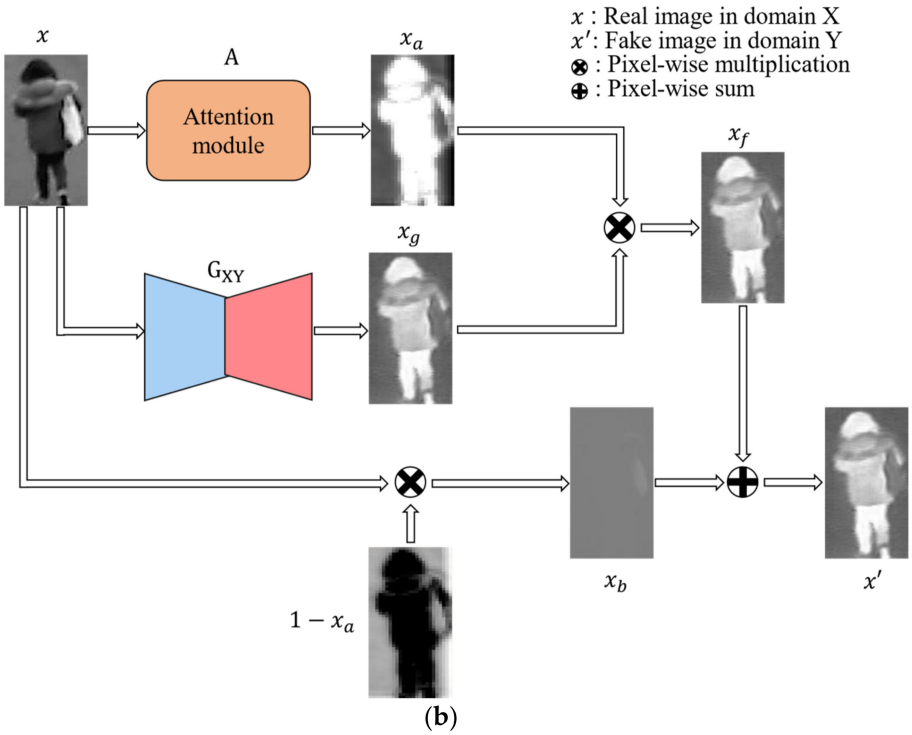


Figure 2. SI-AGAN structure. (**a**) Overall architecture of SI-AGAN. (**b**) Detailed architecture of the black dashed rectangle in (**a**), which includes the attention network.

3.3.1. Attention-Guided Generator Architecture

SI-AGAN consists of two generators and two discriminators for performing style transfer in both directions, from a source domain (X) to a target domain (Y) and from a target domain (Y) to a source domain (X), based on the architecture of AGGAN as a backbone. In this section, the characteristics and structure of an attention-guided generator of SI-AGAN are explained in detail.

Our attention-guided generator, operated as shown in Figure 2b, can be divided into the foreground and background regions. In the foreground region, the input image $x \in X$ of the source domain becomes the input in which x_g is generated through generator G_{XY}. Then, the same input image $x \in X$ becomes the input of an attention module A, and the attention mask x_a becomes the output. An attention mask is obtained from the attention module, as shown in Figure 3, and has a value between [0, 1]. It is trained such that the region requiring attention, or a human region, has a value close to 1. By performing pixel-wise multiplication on the image generated by the generator, the foreground image x_f is generated. In the background region, the value of the previously generated attention mask is reversed to perform pixel-wise multiplication with input image x, thus generating background region image x_b. A final fake image is generated, as expressed in Equation (1), by performing pixel-wise sum for the foreground region image and the background region image.

$$x' = x_a \ G_{XY}(x) + (1 - x_a) \ x \tag{1}$$

Figure 3 and Table 4 show the architecture of the attention module. Figure 4 and Table 5 show the architecture of our attention-guided generator. Figure 5 illustrates the difference between general convolution and depthwise separable convolution used in the proposed SI-AGAN. As shown in Figure 5b, depthwise Separable Convolution has the characteristic that the output values of channels are combined into one. It operates almost similarly to the existing convolution operation, but the number of parameters and the amount of operation reduce. The attention module and generator of the proposed SI-AGAN reduces the computational cost by using a depthwise separable convolution layer. Furthermore, VGG Net-19 based perceptual loss was used in addition to consistency

loss and the least square GAN (LSGAN) loss of the original AGGAN was used for training the generator. The perceptual loss was trained to reduce the difference in feature maps, thus improving training convergence speed as well as the quality of the generated image. A detailed description of the loss is provided in Section 3.3.3.

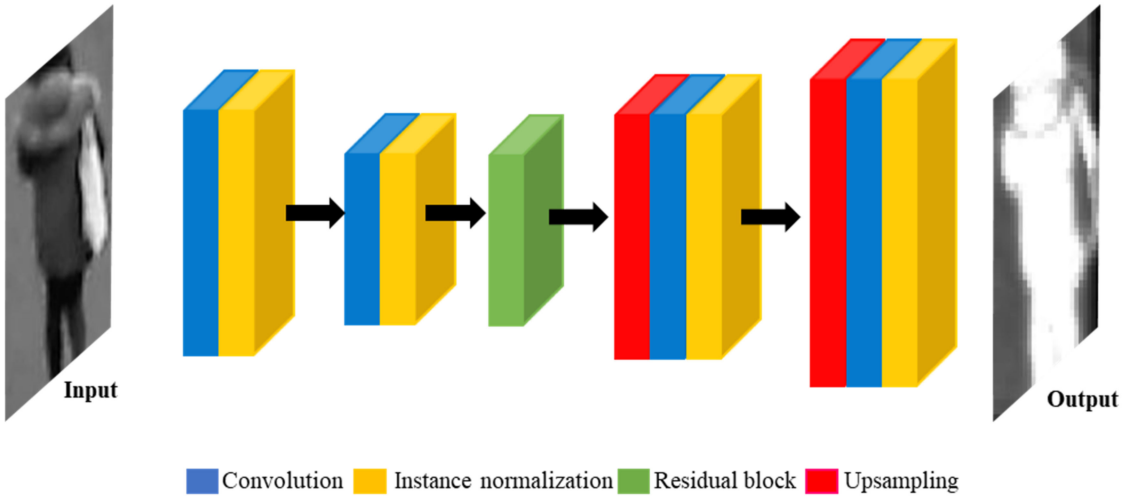


Figure 3. Description of the SI-AGAN attention module.

Table 4. Architecture of the attention module. (Conv, DWConv, IN, and ReLU refer to convolution layer, depthwise separable convolution layer, instance normalization, and rectified linear unit, respectively).

Layer Type		Number of Filters	Size of Feature Map (Width × Height × Channel)	Size of Kernel (Width × Height)	Number of Stride	Number of Padding (Top, Left, Bottom, Right)
Input layer			224 × 100 × 3			
Padding			229 × 105 × 3			(2, 2, 3, 3)
Conv 1 (IN + ReLU)		32	112 × 50 × 32	7 × 7	2 × 2	
Conv 2 (IN + ReLU)		64	56 × 25 × 64	3 × 3	2 × 2	(0, 0, 1, 1)
Residual block	Padding		58 × 27 × 64			(1, 1, 1, 1)
	DWConv (IN + ReLU)	64	56 × 25 × 64	3 × 3	1 × 1	
	Padding		58 × 27 × 64			(1, 1, 1, 1)
	DWConv (IN + ReLU)	64	56 × 25 × 64	3 × 3	1 × 1	
	Add		56 × 25 × 64			
Upsampling layer			112 × 50 × 64			
Conv 3 (IN + ReLU)		64	112 × 50 × 64	3 × 3	1 × 1	(1, 1, 1, 1)
Upsampling layer			224 × 100 × 64			
Conv 4 (IN + ReLU)		32	224 × 100 × 32	3 × 3	1 × 1	(1, 1, 1, 1)
Conv 5 (IN)		1	224 × 100 × 1	7 × 7	1 × 1	(3, 3, 3, 3)
Sigmoid layer			224 × 100 × 1			

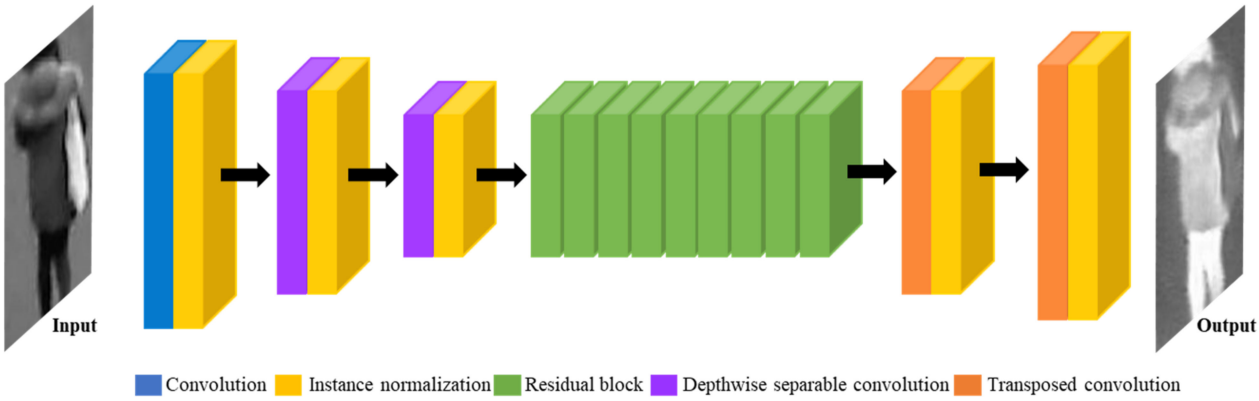


Figure 4. Description of the attention-guided generator of SI-AGAN.

Table 5. Architecture of the attention-guided generator. (Conv, DWConv, IN, ReLU, TransConv, and Tanh each refer to convolution layer, depthwise separable convolution layer, instance normalization, rectified linear unit, transposed convolution, and hyperbolic tangent, respectively; N* indicates a number from 1 to 9).

Layer Type		Number of Filters	Size of Feature Map (Width × Height × Channel)	Size of Kernel (Width × Height)	Number of Stride	Number of Padding (Top, Left, Bottom, Right)
Input layer			224 × 100 × 3			
Padding			230 × 106 × 3			(3, 3, 3, 3)
Conv 1 (IN + ReLU)		32	224 × 100 × 32	7 × 7	1 × 1	
DWConv1 (IN + ReLU)		64	112 × 50 × 64	3 × 3	2 × 2	(0, 0, 1, 1)
DWConv2 (IN + ReLU)		128	56 × 25 × 128	3 × 3	2 × 2	(0, 0, 1, 1)
Residual block N*	Padding		58 × 27 × 128			(1, 1, 1, 1)
	DWConv (IN + ReLU)	128	56 × 25 × 128	3 × 3	1 × 1	
	Padding		58 × 27 × 128			(1, 1, 1, 1)
	DWConv (IN + ReLU)	128	56 × 25 × 128	3 × 3	1 × 1	
	Add		56 × 25 × 128			
TransConv1 (IN + ReLU)		64	112 × 50 × 64	3 × 3	2 × 2	(1, 1, 1, 1)
TransConv2 (IN + ReLU)		32	224 × 100 × 32	3 × 3	2 × 2	(1, 1, 1, 1)
Conv 2 + Tanh		3	224 × 100 × 3	7 × 7	1 × 1	(3, 3, 3, 3)

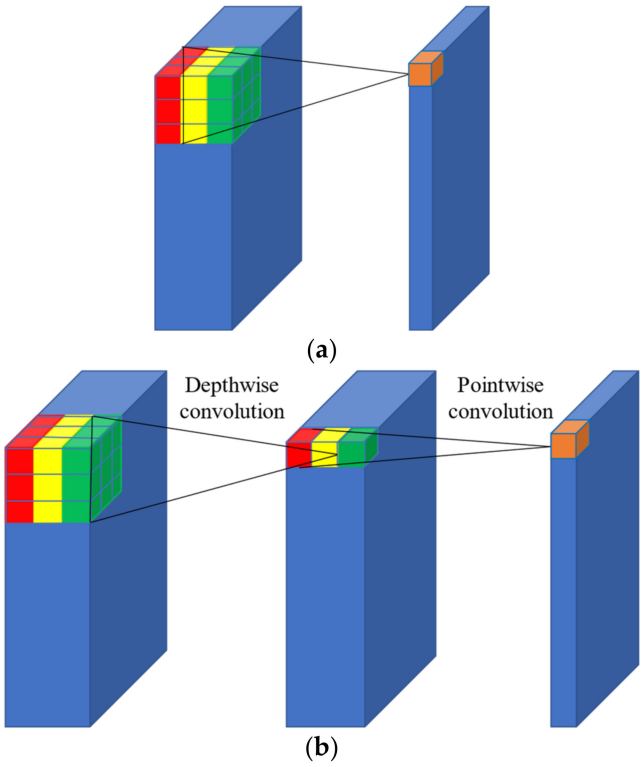


Figure 5. Comparison of convolution. (**a**) General convolution. (**b**) Depthwise separable convolution.

3.3.2. Attention-Guided Discriminator Architecture

In this section, the architecture and characteristics of the attention-guided discriminator of our proposed SI-AGAN are explained in detail. Figure 6 shows the architecture of the SI-AGAN discriminator. The input of a conventional discriminator is a real image, or a fake image generated by a generator. The discriminator is trained to be able to effectively discriminate between real and fake images. A problem associated with this process is that only the foreground region is converted through the attention mask in the attention-guided generator of SI-AGAN, as explained in Section 3.3.1. However, the discriminator distinguishes between real and fake images by considering both foreground and background regions. Hence, generator performance is affected as the discriminator becomes less effective as the training proceeds. To solve this problem, a switch parameter s is set in our attention-guided discriminator to use a real image before reaching the epoch corresponding to the switch parameter, and then the real image considering the attention module is used as the input after reaching the respective epoch. This is represented as a blue dash-single dotted line in Figure 2a, in which the input of the discriminator is determined by the switch parameter. For the real image considering the attention module, the $a \in x_a$ (attention mask) value generated in the attention module, as expressed in Equation (2), is updated to 1 when higher than or equal to the mask threshold parameter t (set to 0.1 in this study) or to 0 otherwise, and then pixel-wise multiplication with the real image is performed.

$$a_{new} = \begin{cases} 1 \; if \; a \geq t \\ 0 \; if \; a < t \end{cases} \tag{2}$$

Figure 6 and Table 6 show the architecture of the attention-guided discriminator. The discriminator of the proposed SI-AGAN reduced the computational cost by using the depthwise separable convolution layer.

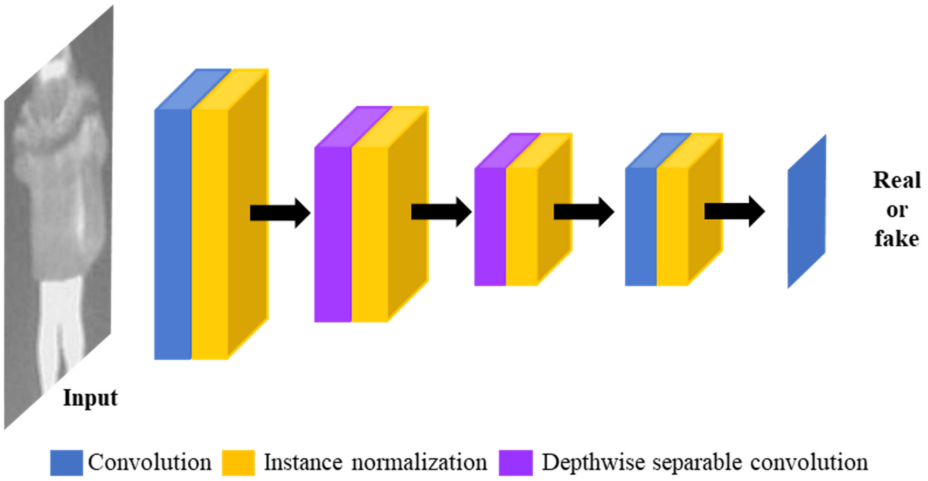


Figure 6. Description of the SI-AGAN discriminator.

Table 6. Architecture of attention-guided discriminator. (Conv, DWConv, IN, and LReLU refer to convolution layer, depthwise separable convolution layer, instance normalization, and leaky rectified linear unit, respectively; N* indicates a number from 1 to 9; for IN*, IN is executed until reaching the epoch corresponding the pre-determined switch parameter, and IN is not executed afterward).

Layer Type	Number of Filters	Size of Feature Map (Width × Height × Channel)	Size of Kernel (Width × Height)	Number of Stride	Number of Padding (Top, Left, Bottom, Right)
Input layer		224 × 100 × 3			
Padding		228 × 104 × 3			(2, 2, 2, 2)
Conv 1 (IN* + LReLU)	64	113 × 51 × 64	4 × 4	2 × 2	
Padding		116 × 54 × 64			(1, 1, 2, 2)
DWConv 1 (IN* + LReLU)	128	57 × 26 × 128	4 × 4	2 × 2	
Padding		60 × 30 × 128			(1, 2, 2, 2)
DWConv 2 (IN* + LReLU)	256	29 × 14 × 256	4 × 4	2 × 2	
Padding		33 × 18 × 256			(2, 2, 2, 2)
Conv 2 (IN* + LReLU)	512	30 × 15 × 512	4 × 4	1 × 1	
Padding		34 × 19 × 512			(2, 2, 2, 2)
Output layer (Conv)	1	31 × 16 × 1	4 × 4	1 × 1	

3.3.3. Loss Function of SI-AGAN

In the original AGGAN, LSGAN loss was used to generate a sharper image; the cycle-consistency loss proposed in the cycle-consistent adversarial networks (CycleGAN) [38] was also used to prevent the identity of the input image from being considerably modified. In our SI-AGAN, the perceptual loss was additionally used for the losses of the original AGGAN to improve the quality of the generated person image.

First, LSGAN expressed in Equation (3) below was introduced for adversarial training between the generator and discriminator of SI-AGAN. Here, D_{XY} is the attention-guided discriminator from the source domain (X) toward the target domain (Y), G_{XY} is the attention-

guided generator from the source domain toward the target domain, and A_X is the attention module in the source domain.

$$\mathcal{L}^x{}_{\mathrm{GAN}}(G_{XY}, A_X, D_{XY}) = \mathbb{E}_{\mathrm{x}\sim \mathrm{p}_{\mathrm{data}}(\mathrm{x})}\left[D_X(G_{XY}(x))^2\right] + \mathbb{E}_{\mathrm{y}\sim \mathrm{p}_{\mathrm{data}}(\mathrm{y})}\left[(D_X(y)-1)^2\right] \quad (3)$$

Second, the cycle-consistency loss expressed in Equation (4) was introduced to prevent the identity of the input image of SI-AGAN from being modified substantially. Here, x is the input image $x \in X$ of the source domain. The consistency loss is trained so that x becomes less different from the image obtained through inverse mapping, or the reconstructed image in Figure 2a, thus preventing the input image's identity from being significantly modified.

$$\mathcal{L}^x{}_{\mathrm{CYC}}(G_{XY}, A_X) = \mathbb{E}_{\mathrm{x}\sim \mathrm{p}_{\mathrm{data}}(\mathrm{x})}||x - G_{YX}(G_{XY}(x))||_1 \quad (4)$$

The cycle-consistency loss is trained to simply reduce the pixel difference between the input image and the reconstructed image. Here, our SI-AGAN reduced the difference in feature maps through VGG-based perceptual loss of the input image and reconstructed image. In [11], the difference in feature maps of visible-light images and IR images was shown to illustrate that the concentrated regions vary on feature maps. Accordingly, VGG-based perceptual loss expressed in Equation (5), which is trained to reduce the difference in the feature step between the reconstructed image and the input image x of source domain that are on the same domain, was added in this paper for generating a syn-IR image, which is similar to an IR image. Here, $\varnothing_{\mathrm{i}}(x)$ is the feature map of x extracted from the i-th layer of VGG Net-19, while H_i, W_i, and C_i refer to the height, width, and the channel of a feature map for x extracted from the i-th layer, respectively.

$$\mathcal{L}^x{}_{\mathrm{PER}}(G_{XY}, A_X) = \frac{1}{H_i W_i C_i}\sum_{\mathrm{h}=1}^{H_i}\sum_{\mathrm{w}=1}^{W_i}\sum_{\mathrm{c}=1}^{C_{\mathrm{i}}}\left(\varnothing_{\mathrm{i}}(x)_{\mathrm{h,\ w,c}} - \varnothing_{\mathrm{i}}(G_{YX}(G_{XY}(x)))_{\mathrm{h,w,c}}\right)^2) \quad (5)$$

The losses explained thus far are for times when training is performed from the source domain (X) toward the target domain (Y). The same losses are applied in the inverse direction from the target domain toward the source domain. Accordingly, the final SI-AGAN loss is as expressed in Equation (6). Here, λ_{CYC} d λ_{PER} are loss hyper-parameters for our experiment.

$$\mathcal{L}(G_{XY}, G_{YX}, A_X, A_Y, D_{XY}, D_{YX}) = \mathcal{L}^x{}_{\mathrm{GAN}} + \mathcal{L}^y{}_{\mathrm{GAN}} + \lambda_{\mathrm{CYC}}(\mathcal{L}^x{}_{\mathrm{CYC}} + \mathcal{L}^y{}_{\mathrm{CYC}}) + \lambda_{\mathrm{PER}}(\mathcal{L}^x{}_{\mathrm{PER}} + \mathcal{L}^y{}_{\mathrm{PER}}) \quad (6)$$

3.3.4. Differences between the Proposed SI-AGAN and Original AGGAN

In this section, the differences between the proposed SI-AGAN and the original AGGAN are summarized:

- In the original AGGAN, a square image is used as an input. However, body shapes and body proportions of males and females provide critical information regarding gender recognition. Therefore, the proposed SI-AGAN was trained using vertically long rectangular input images instead of square images.
- To reduce the computational cost, certain convolutional layers of the original AGGAN were revised to depthwise separable convolutional layers in the SI-AGAN. The revised convolutional layers are the entire convolutional layers of the residual blocks in the attention module, second and third convolutional layers of the attention-guided generator, entire convolutional layers of the residual blocks in the attention-guided generator, and second and third convolutional layers of the attention-guided discriminator.
- Finally, VGG Net-19-based perceptual loss was applied between the input image and the reconstructed image in SI-AGAN. While training the SI-AGAN, pixels of the images on the same domain and the quality of the image generated by considering

the difference in feature maps were improved, thus enhancing the gender recognition performance.

3.4. CNN and Score-Level Fusion for Gender Recognition

The reconstructed visible-light image obtained through two-step image reconstruction and the syn-IR image generated through SI-AGAN were used as the input of ResNet-101 to obtain the scores, which were then applied with SVM-based score-level fusion to finally perform gender recognition. The existing ResNet-101 was trained with a train from scratch method using the training data of this study. ResNet-101 has a total of five stages, in which stages 2–5 consist of convolutional blocks and identity blocks [30]. Once the five stages are completed, a fully connected layer is configured after average pooling, and gender recognition finally proceeds through the Softmax layer. The reconstructed visible-light image and the syn-IR image generated by SI-AGAN are applied to the ResNet-101 to obtain scores from the fully connected layer and perform score-level fusion.

The score obtained from the reconstructed visible-light image and the score obtained from the syn-IR image undergo normalization first for the stable performance of SVM. For finding the optimal performance in this study, six normalization methods (standard scaler, min-max scaler, robust scaler, normalizer scaler, quantile transformer, power transformer) were compared, whereas SVM was compared with the linear kernel, radial basis function (RBF) kernel, polynomial kernel, and sigmoid kernel. Each kernel function for mapping the vector of a low-dimensional space to the vector of a high-dimensional space can be expressed as in Equations (7)–(10). Normalization and kernel function proceeded using the optimal value found in the training data.

$$K(s_i, s_j) = s_i^T s_j \text{ (Linear kernel)} \tag{7}$$

$$K(s_i + s_j) = exp\left\{-\frac{\| s_i - s_j \|_2^2}{2\sigma^2}\right\},\ \sigma \neq 0 \text{ (RBF kernel)} \tag{8}$$

$$K(s_i, s_j) = \left(s_i^T s_j + c\right)^d,\ c > 0) \text{ (Polynomial kernel)} \tag{9}$$

$$K(s_i, s_j) = \tanh\left\{a\left(s_i^T s_j\right) + b\right\}',\ a, b \geq 0 \text{ (Sigmoid kernel)} \tag{10}$$

For the SYSU-MM01 database with a large number of images, computational time was measured in the desktop environment explained in Section 4.2. There are a total of 9819 training images, and the computational time of each kernel of Equations (7)–(10) with 9819 images was measured to be 4.9, 305.2, 8.9, and 11.9 ms, respectively. The computational time with a total of 3727 test images were measured to be 2.9, 117.2, 3, and 5.9 ms, respectively, for each kernel of Equations (7)–(10). The processing time per image shows 0.03 ms for both the training and test images based on the RBF kernel of Equation (8), which takes the longest processing time. In this paper, SVM shows fast processing time by using two scores extracted from each CNN step as input for one image.

The subjects can be finally classified into male and female based on the threshold of the score obtained through SVM. During classification, incorrect classification of a male image as a female image is a Type I error, while incorrect classification of a female image as a male image is a Type II error. Type I and Type II errors have a tradeoff relationship. The value when Type I and Type II errors match is defined as the equal error rate (EER). In this study, the point of obtaining the EER was used as the threshold for classifying the gender.

4. Experimental Results

4.1. Experimental Database and Environment

For the first experiment, the RegDB database [39], which is an open database, was used for gender recognition. The human images in the RegDB database were captured in uncontrolled environments using one visible-light camera and one thermal camera. The

RegDB database has images of moving persons taken outdoors in uncontrolled environments, as shown in Figure 7a; thus, low-resolution images with severe blur and noise were captured. Moreover, images were captured using visible-light and thermal cameras simultaneously; thus, the images were paired for the same pose. The RegDB database consists of 4120 visible-light and thermal images and 412 human classes. Fivefold cross-validation was applied in the experiment as the total number of images in the database is small in which the classes of different persons were configured for fivefold cross-validation (open world setting). During the first fold, 3310 images among 4120 images were applied with data augmentation based on translation and cropping to obtain a total of 74,820 images, which were then used as the training set, as shown in Table 7, while 810 images were used as the test dataset.

Figure 7. Example of the experiment database. (**a**) RegDB database and (**b**) SYSU-MM01 database.

Table 7. Descriptions of the experimental database.

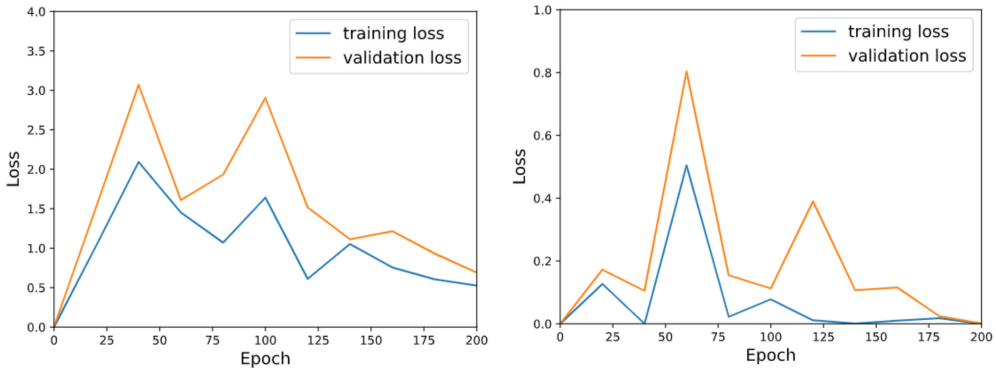
Database	**RegDB**		**SYSU-MM01**		
Subset	**Training**	**Test**	**Training**	**Validation**	**Testing**
Number of people (male/female)	331 (204/127)	81 (50/31)	295 (154/141)	96 (58/38)	99 (63/36)
Number of images (male/female)	3310 (2040/1270)	810 (500/310)	9819 (5190/4629)	1949 (1259/690)	3727 (2283/1444)

For the second experiment, the SYSU-MM01 database [40], which is also an open database, was used for gender recognition. The human images in the SYSU-MM01 database were captured in both indoor and outdoor environments, as shown in Figure 7b, using four visible-light cameras and two NIR cameras. Visible-light images were captured in the daytime, while NIR images were captured in the nighttime; thus, the database consists of unpaired images as the person of the same class was captured at different times. Also, a person of the same class may have different images depending on clothes, bags, or accessories. The original SYSU-MM01 database consists of 287,628 visible-light images, 15,792 NIR images, and 691 human classes. In this study, the numbers of visible-light images and NIR images were set to be identical for the same class. If the number of visible-light images is greater in the same class, the images that can be easily used to recognize gender because they were captured from a relatively close distance have been excluded from the experiment. The same process was applied for the opposite case. A total of

15,495 images of the SYSU-MM01 database were used for gender recognition in this study; as shown in Table 7, 9819 images were used for the training dataset, 1949 images were used for the validation dataset, and 3727 images were used for the testing dataset, as specified by the database provider. Training, validation, and testing datasets are configured so that the classes do not overlap (open-world setting).

4.2. Training of SI-AGAN and CNN Models

The adaptive moment estimation (ADAM) [41] was used as an optimizer for training SI-AGAN. The initial learning rate was set to 0.002, beta 1–0.5, and beta 2–0.999. The RegDB database was trained for a total of 200 epochs in which the learning rate was maintained at 0.002 until 100 epochs then gradually became 0 at 200 epochs. When classification was performed for real images in the attention-guided discriminator of SI-AGAN, the switch parameter determined when to apply the attention module for the real images. The switch parameter was set to 30 in this study where classification was performed with real images of the discriminator before 30 epochs, and then classification was performed with the images applied with the attention module after 30 epochs. The SYSU-MM01 database is an unpaired dataset that is difficult to be trained at first; thus, the model in which the RegDB database was trained was fine-tuned. Other parameters matched with the RegDB database, and the switch parameter was set to 0 because training values were initially available from fine-tuning. Figure 8 shows the training and validation loss curves of attention-guided generator and attention-guided discriminator of SI-AGAN. In GAN, a generator is usually more complicated than a discriminator because the generator creates an image. Therefore, the loss value of discriminator tends to be lower than that of the generator as shown in Figure 8 because the discriminator simply performs binary classification [34]. The reason why there are oscillates in the loss graphs of Figure 8 is as follows. We used a switch parameter of 30 epochs for training, which means our attention module is operated at the first time after 30 epochs, which causes oscillates in the loss graphs of Figure 8a. Also, before 100 epochs, the learning rate is fixed, but we made the learning rate go down after 100 epochs, which causes another oscillates in the loss graphs of Figure 8a,b. Nevertheless, as the learning rate decreases afterward, the training loss graphs converge stably as shown in Figure 8.



(**a**) (**b**)

Figure 8. Training and validation loss graphs. (**a**) Training and validation loss graph of attention-guided generator of SI-AGAN and (**b**) training and validation loss graph of attention-guided discriminator of SI-AGAN.

ResNet-101 was used as the CNN model for performing gender recognition in this study. Stochastic gradient descent (SGD) [42] was used as an optimizer for training ResNet-101. The initial learning rate was set to 0.01, momentum to 0.9, and weight decay to 0.0001. The learning rate was optimized by multiplying with a gamma value every 10 epochs based on the stepped policy. ResNet-101 trained the image applied with two-step CNN-based reconstruction and the syn-IR image generated through SI-AGAN. Both the RegDB database and the SYSU-MM01 database used in the experiment were trained with the same

parameters. Figure 9 shows the training and validation loss and accuracy of ResNet-101. The loss converged to a low value as the training epochs increased, whereas the training accuracy converged to nearly 100%. Thus, ResNet-101 was considered stably trained. As shown in the validation loss and accuracy graphs in Figure 9, ResNet-101 was also not overfitted by the training data.

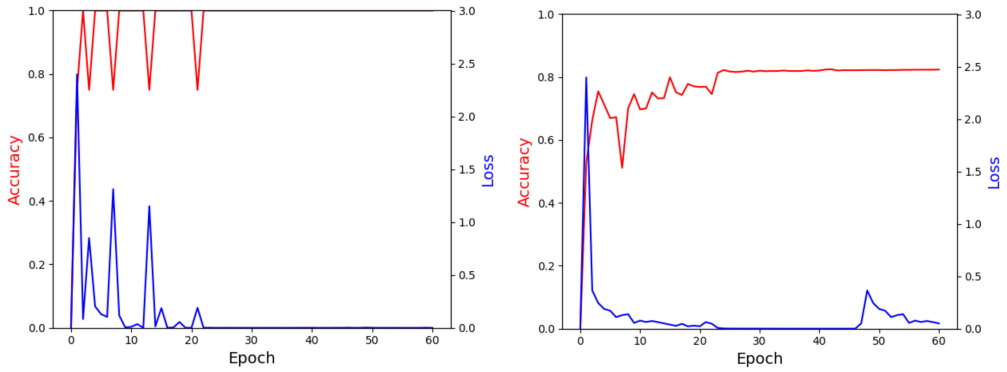


Figure 9. Training and validation loss and accuracy graphs of ResNet-101. (**a**) Training loss and accuracy graph. (**b**) Validation loss and accuracy graph.

The proposed algorithm was implemented using MatConvNet (version 1.0-beta 25) [43], Caffe framework (version 1.0.0) [25], and TensorFlow-GPU 1.12.0 [44]. The experiment was conducted using a PC equipped with Intel® Core™ i7-7700 CPU @ 3.6 GHz (4 cores) with 32 GB of main memory, and NVIDIA GeForce GTX 1070 Ti (2432 compute unified device architecture (CUDA) cores) with a graphics memory of 8 GB (NVIDIA, Santa Clara, CA, USA) [45].

4.3. Testing of SI-AGAN and CNN Models with RegDB

4.3.1. Ablation Studies

As the first ablation study, we evaluated the performance of two-step CNN-based reconstruction. As shown in Table 8 and Figure 10, the recognition performance was degraded when an IRCNN was applied to visible-light images, whereas the best performance was exhibited when IRCNN- and VDSR-based two-step reconstruction methods were applied.

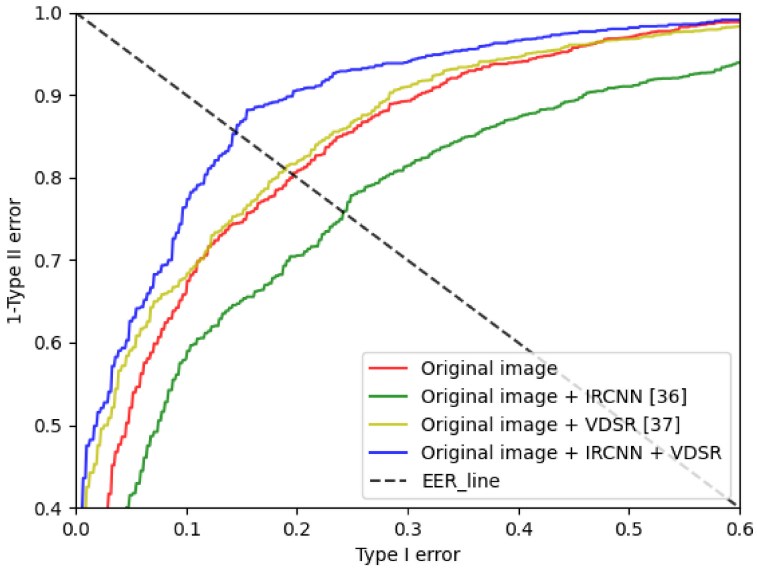


Figure 10. ROC curves of gender recognition accuracies using visible-light images.

Table 8. Comparisons of gender recognition accuracies using reconstructed visible-light images.

Methods	5-Fold Cross Validation	EER (%)	
		1~5 Fold	Average
Original image	1 fold	16.78	19.54
	2 fold	22.32	
	3 fold	23.21	
	4 fold	19.01	
	5 fold	16.42	
Original image + IRCNN [36]	1 fold	24.81	24.25
	2 fold	23.73	
	3 fold	25.96	
	4 fold	22.59	
	5 fold	24.19	
Original image + VDSR [37]	1 fold	13.73	18.75
	2 fold	19.57	
	3 fold	20.46	
	4 fold	20.98	
	5 fold	19.01	
Original image + IRCNN + VDSR (proposed method)	1 fold	14.09	15.07
	2 fold	18.65	
	3 fold	14.09	
	4 fold	14.45	
	5 fold	14.09	

In the second ablation study, the performance was compared to that of the proposed SI-AGAN with or without perceptual loss and depthwise separable convolution. As shown in Table 9 and Figure 11, SI-AGAN with perceptual loss and depthwise separable convolution exhibited the best gender recognition performance.

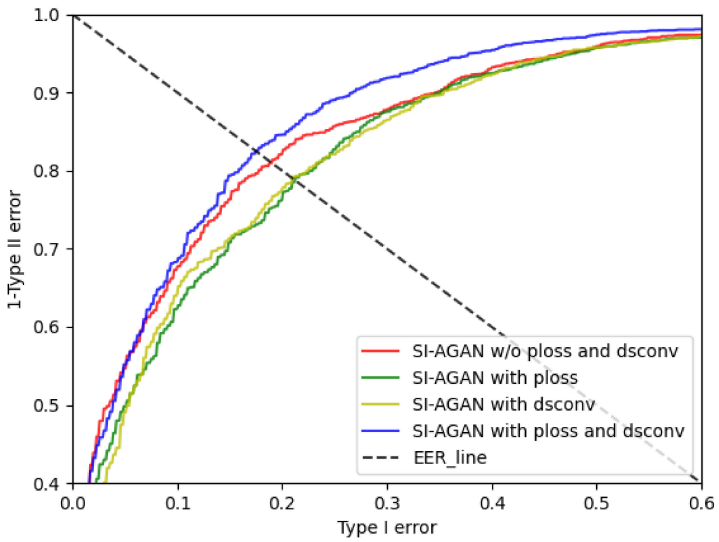


Figure 11. ROC curves of gender recognition accuracies of SI-AGAN with or without perceptual loss and depthwise separable convolution. (w/o refers to without, ploss refers to perceptual loss, dsconv refers to depthwise separable convolution).

Table 9. Comparison of the gender recognition accuracies of SI-AGAN with or without perceptual loss and depthwise separable convolution.

Methods	5-Fold Cross Validation	EER (%)	
		1~5 Fold	Average
SI-AGAN without perceptual loss and depthwise separable convolution	1 fold	14.55	19.01
	2 fold	16.62	
	3 fold	24.61	
	4 fold	19.21	
	5 fold	20.10	
SI-AGAN with perceptual loss	1 fold	24.45	20.99
	2 fold	18.39	
	3 fold	25.08	
	4 fold	21.86	
	5 fold	15.18	
SI-AGAN with depthwise separable convolution	1 fold	18.65	20.74
	2 fold	17.67	
	3 fold	23.83	
	4 fold	18.49	
	5 fold	25.08	
SI-AGAN with perceptual loss and depthwise separable convolution (proposed method)	1 fold	18.03	17.65
	2 fold	14.45	
	3 fold	19.37	
	4 fold	19.27	
	5 fold	17.14	

4.3.2. Comparative Experiments of SI-AGAN with the State-of-the-Art Methods for Style Transfer

This section describes the comparative experiments of SI-AGAN with state-of-the-art methods for style transfer. For state-of-the-art methods for style transfer, CycleGAN [38], ThermalGAN [46], and AGGAN [47] were used. Furthermore, the recognition network was fixed to be ResNet-101 for fair comparisons; the generated image was trained using the train from scratch method in the same environment proposed in Section 4.2.

Table 10 and Figure 12 show the performance results of recognizing IR images generated by various GAN models measured through ResNet-101. Our proposed method, SI-AGAN, exhibited better performance than conventional GAN model in which CycleGAN and ThermalGAN had poorer performance than the case in which the original visible-light image was used, as shown in Table 8. Our proposed SI-AGAN trains human images through the attention module and generates images with more focus on the human region, thus exhibiting outstanding performance. Figure 13 shows the examples of the generated syn-IR image. Relatively clear visible-light images adequately generate IR images in all GAN models. For visible-light images with severe noise or blur, however, the quality of the IR images generated by the conventional GAN models was significantly reduced. Certain images with bags or accessories were also not generated properly in the GAN models. Therefore, the syn-IR images generated by the proposed SI-AGAN have excellent gender recognition performance as well as visibility of the generated images.

Table 10. Gender recognition accuracies of various style transfer methods.

Methods	5-Fold Cross Validation	EER (%)	
		1~5 Fold	Average
CycleGAN [38]	1 fold	21.18	19.37
	2 fold	17.67	
	3 fold	20.62	
	4 fold	20.52	
	5 fold	16.88	
ThermalGAN [46]	1 fold	26.42	24.57
	2 fold	28.61	
	3 fold	20.16	
	4 fold	27.20	
	5 fold	20.46	
AGGAN [47]	1 fold	14.55	19.01
	2 fold	16.62	
	3 fold	24.61	
	4 fold	19.21	
	5 fold	20.10	
SI-AGAN	1 fold	18.03	17.65
	2 fold	14.45	
	3 fold	19.37	
	4 fold	19.27	
	5 fold	17.14	

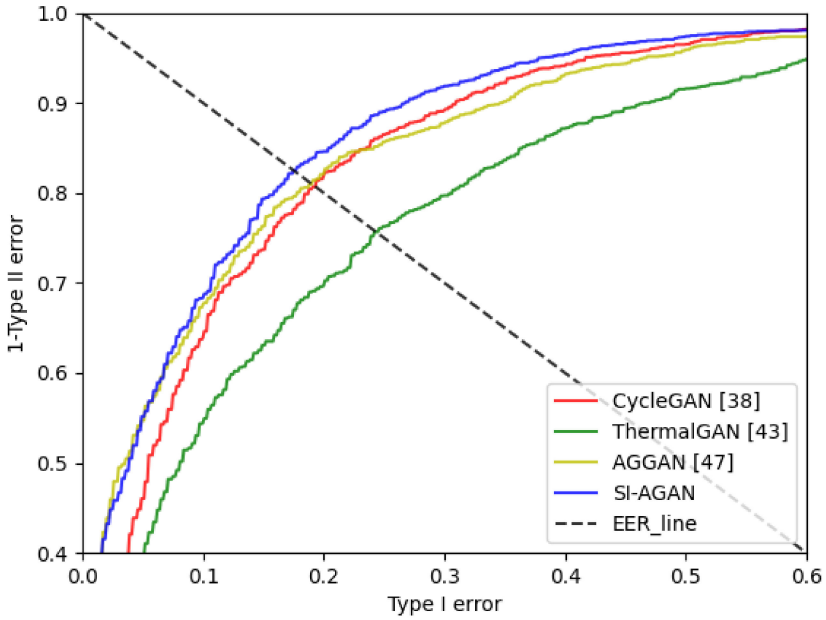


Figure 12. ROC curves of gender recognition accuracies with various style transfer methods.

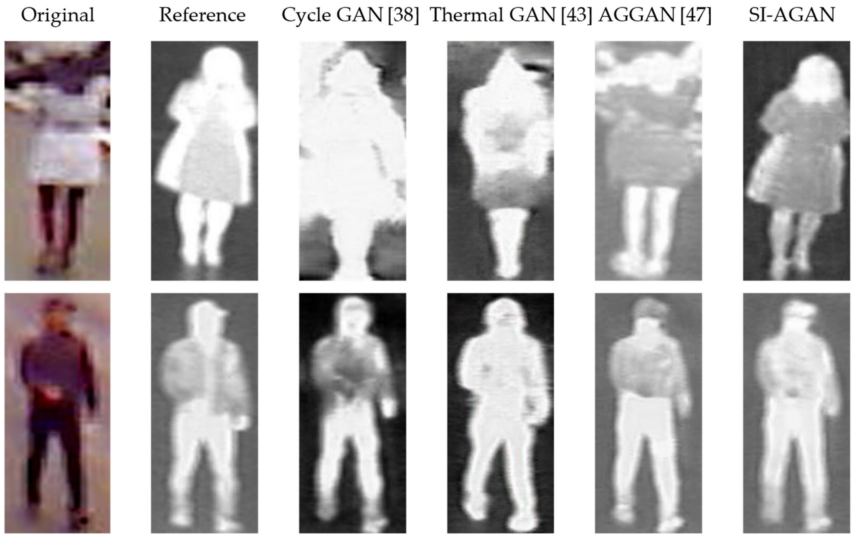


Figure 13. Comparisons of synthesized images according to different GAN models.

4.3.3. Recognition Accuracies Based on Score-Level Fusion and Comparisons with State-of-the-Art Methods

In this section, the final gender recognition is compared by conducting SVM-based score-level fusion for the syn-IR images generated by the GAN models and the reconstructed visible-light images. Table 11 and Figure 14 show the comparisons of final gender recognition performance, where SVM-based score-level fusion is applied to the syn-IR images generated by various GAN models and the visible-light image is applied with IRCNN and VDSR. Our proposed method was found to be superior in terms of single performance of syn-IR images and the combined performances. Figure 15 shows the Type I and Type II errors of the proposed method and correct cases. As shown in the images of Type I and Type II errors, recognition is rather unsuccessful if the original image has severe noise or blur, which hinder gender recognition, or the image is distorted severely during the two-step image reconstruction process. Correct cases were classified correctly even when it was difficult to perform gender classification using the image.

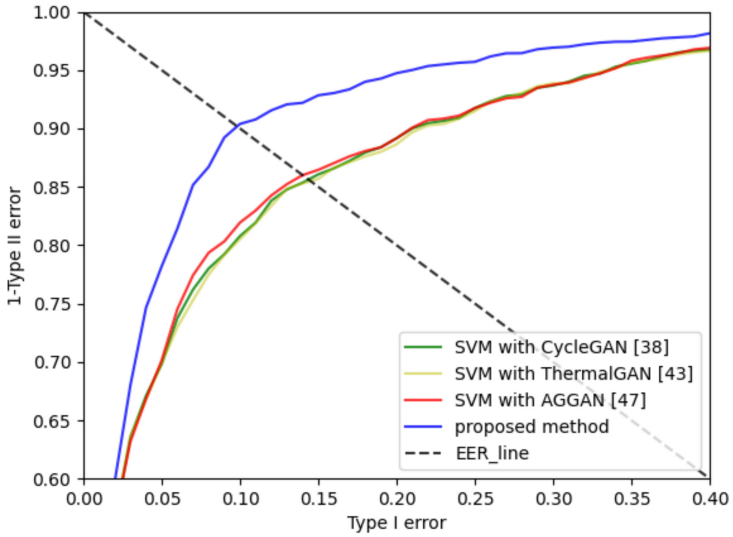


Figure 14. ROC curves of gender recognition accuracies using score-level fusion (SVM with CycleGAN [38] refers to Visible-light image (+IRCNN+VDSR) + syn-IR image (CycleGAN) [38], SVM with ThermalGAN [46] refers to Visible-light image (+IRCNN+VDSR) + syn-IR image (ThermalGAN) [46], SVM with AGGAN [47] refers to Visible-light image(+IRCNN+VDSR) + syn-IR image (AGGAN) [47]).

Table 11. Comparison of gender recognition accuracies based on score-level fusion.

Methods	5-Fold Cross Validation	EER (%)	
		1~5 Fold	Average
Visible-light image (+IRCNN+VDSR) + syn-IR image (CycleGAN) [38]	1 fold	14.31	13.63
	2 fold	19.74	
	3 fold	18.35	
	4 fold	14.77	
	5 fold	1.00	
Visible-light image (+IRCNN+VDSR) + syn-IR image (ThermalGAN) [46]	1 fold	14.47	13.57
	2 fold	18.60	
	3 fold	19.20	
	4 fold	14.49	
	5 fold	1.10	
Visible-light image (+IRCNN+VDSR) + syn-IR image (AGGAN) [47]	1 fold	13.61	12.92
	2 fold	18.03	
	3 fold	17.39	
	4 fold	14.43	
	5 fold	1.16	
Visible-light image (+IRCNN+VDSR) + syn-IR image (SI-AGAN) (proposed method)	1 fold	9.94	9.05
	2 fold	9.16	
	3 fold	13.51	
	4 fold	10.92	
	5 fold	1.75	

Moreover, our proposed method was compared with previous methods in which visible-light image and IR image are combined. As shown in Table 12, our proposed method exhibited better performances than previous methods. In previous studies, various methods have been researched to extract important features of the gender of a person from visible-light and IR images. HOG features showed poor performance in the initial experiment, but the performance improved by applying the weighted HOG, which can focus more on the human region using the characteristics of IR images. The possibility of utilizing the handcrafted features was proven through research on deep features using CNNs as the technologies related to CNNs continue to advance. Furthermore, visible-light images were improved through two-step image reconstruction in [11] methods, while gender recognition performance was improved through SI-AGAN, which was used to generate syn-IR images in our proposed method. Both visible-light images and IR images are required for testing in previous methods, but gender recognition can be performed only with visible-light images in our proposed method.

We performed *t*-test [48] and measured Cohen's d-value [49] between proposed method and the second-best method in Tables 11 and 12 for the statistical test. Cohen's d-value around 0.2 means a small effective size, 0.5 means a medium effective size, and 0.8 means a large effective size. As shown in Figure 16a, we measured the *p*-values of the second-best method and our proposed method in Table 11. The *p*-value of result was 0.318, which means a 68% confidence level, and Cohen's d-value was 0.67 (medium effective size). As shown in Figure 16b, we measured the *p*-values of the second-best method and our proposed method in Table 12. The *p*-value of result was 0.423, which means a 57% confidence level, and Cohen's d-value was 0.53 (medium effective size).

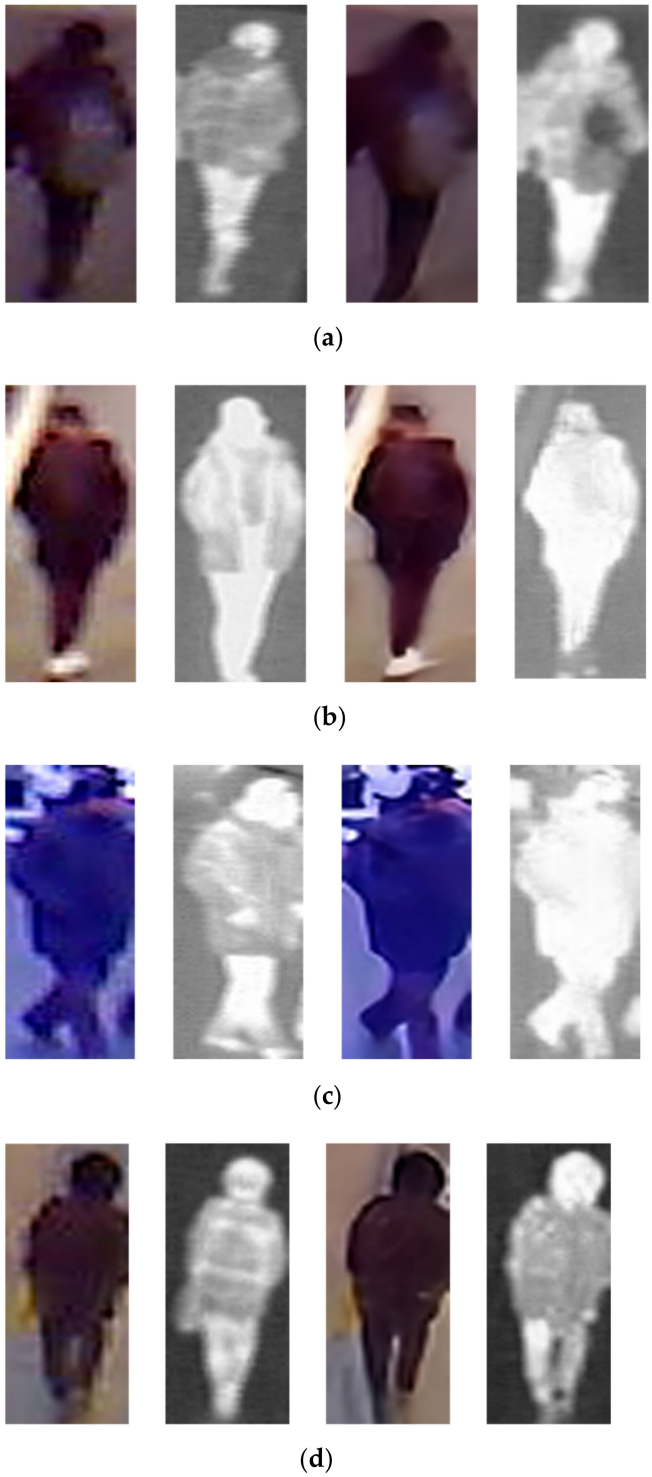
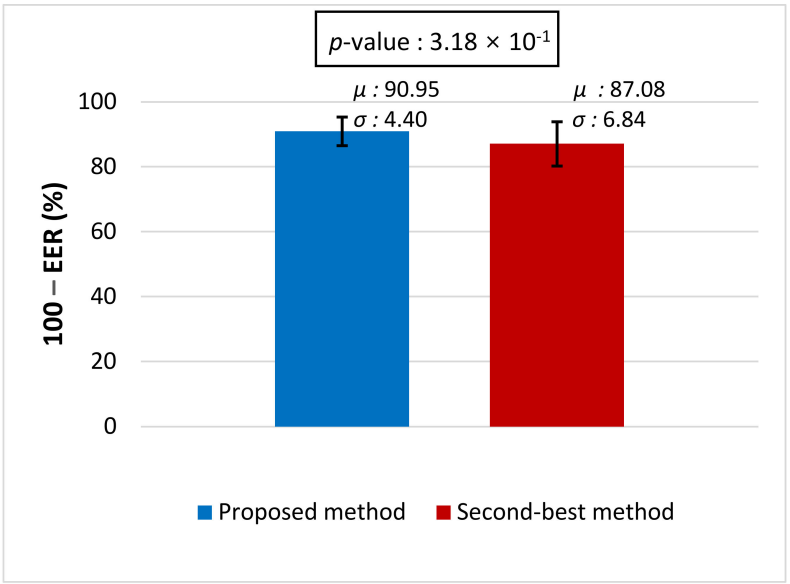


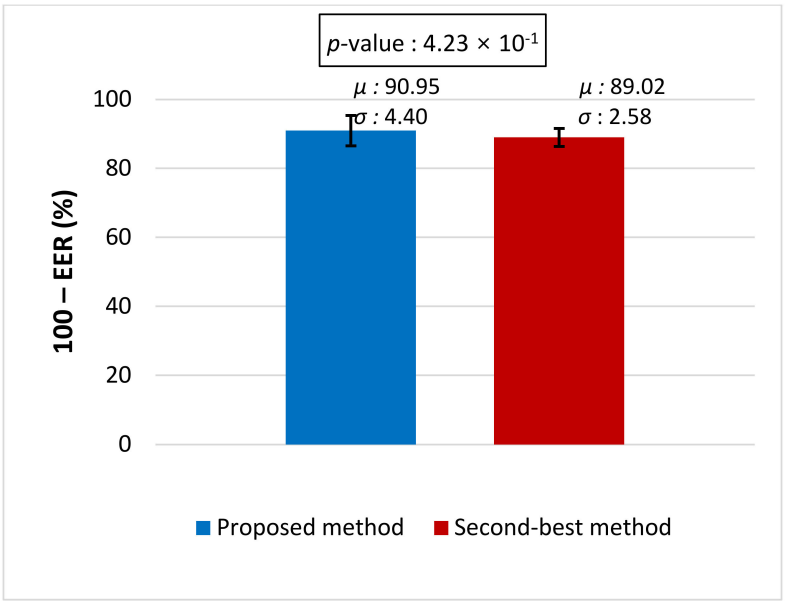
Figure 15. Cases of Type I error, Type II error, correct recognition. (**a**) Cases of Type I error, (**b**) cases of Type II error, (**c**) correct recognition cases (male), (**d**) correct recognition cases (female). In (**a**–**d**), from left to right are original visible-light image, original IR image, reconstructed visible-light image, and syn-IR image.

Table 12. Comparison of gender recognition accuracies with our method and previous methods.

Methods	EER (%)
Visible-light image + IR image using HOG feature [4,22]	16.28
Visible-light image + IR image using weighted HOG feature [23]	13.06
Visible-light image + IR image using AlexNet [9,33]	11.71
Visible-light image (+IRCNN+VDSR) + IR image using ResNet-101 [11]	10.98
Visible-light image (+IRCNN+VDSR) + syn-IR image (SI-AGAN) using ResNet-101 (proposed method)	9.05



(**a**)



(**b**)

Figure 16. *T*-test result between our proposed method and the second-best method. (**a**) Comparison between proposed method and Visible-light image(+IRCNN+VDSR) + syn-IR image (AGGAN) [47] in Table 11. (**b**) Comparison between proposed method and Visible-light image (+IRCNN+VDSR) + IR image using ResNet-101 [11] in Table 12.

4.4. Testing of SI-AGAN and CNN Models with SYSU-MM01

In this section, an experiment was conducted using the SYSU-MM01 database. The reconstruction performance of the visible-light image is explained first, and then the performance of the methods for generating syn-IR images is explained afterwards. Finally, the final performance where two images are applied with SVM-based score-level fusion is explained.

4.4.1. Ablation Studies

As the first ablation study, the performance of applying CNN-based reconstruction to a visible-light image was compared with original image. As shown in Table 13 and Figure 17, higher recognition performance was exhibited when two-step image reconstruction was applied

Table 13. Comparisons of gender recognition accuracies using the reconstructed visible-light images.

Methods	EER (%)
Original image	16.24
Original image + IRCNN + VDSR (proposed method)	14.90

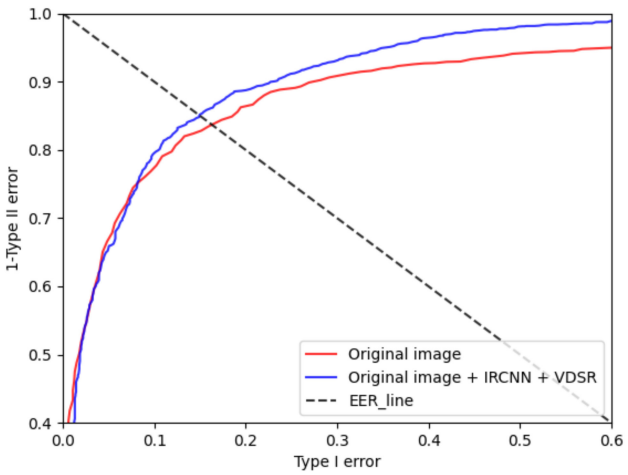


Figure 17. ROC curves of gender recognition accuracies with visible-light images.

As in the second ablation study, Table 14 and Figure 18 show the comparisons of gender recognition accuracies of SI-AGAN with or without perceptual loss and depthwise separable convolution. As shown in Table 14 and Figure 18, SI-AGAN with perceptual loss and depthwise separable convolution exhibited higher recognition performance than SI-AGAN without perceptual loss and depthwise separable convolution.

Table 14. Comparisons of gender recognition accuracies using the reconstructed visible-light images.

Methods	EER (%)
SI-AGAN without perceptual loss and depthwise separable convolution	39.13
SI-AGAN with perceptual loss and depthwise separable convolution (proposed method)	25.21

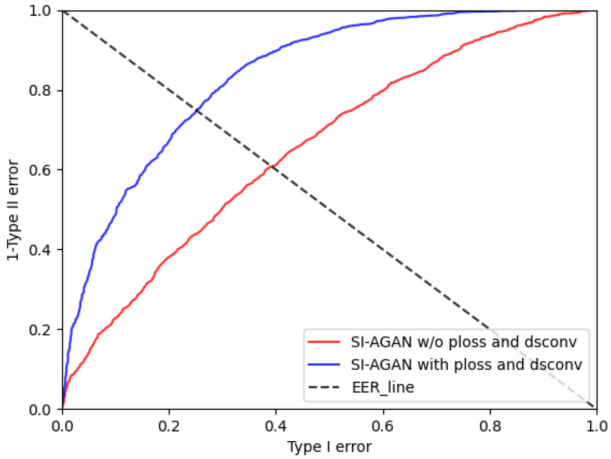


Figure 18. ROC curves of gender recognition accuracies with SI-AGAN with or without perceptual loss and depthwise separable convolution. (w/o refers to without, ploss refers to perceptual loss, dsconv refers to depthwise separable convolution).

4.4.2. Recognition Accuracies Based on Score-Level Fusion and Comparisons with State-of-the-Art Methods

Table 15 and Figure 19 show the comparisons of the final gender recognition where SVM-based score-level fusion is applied to the syn-IR images generated by various GAN models and the visible-light image applied with IRCNN and VDSR. Our proposed method is superior in the single performance of syn-IR images as well as in the combined performances. Figure 20 shows the Type I and Type II errors of the proposed method and correct recognition cases. As shown in the images of Type I and Type II errors, recognition is rather unsuccessful if the original image has severe noise or blur to hinder gender recognition or the image is distorted severely during the two-step image reconstruction process. Correct recognition cases were classified correctly even when the image made it difficult to do so.

Table 15. Comparison of gender recognition accuracies based on score-level fusion.

Methods	EER (%)
Visible-light image (+IRCNN+VDSR) + syn-IR image (AGGAN)	17.60
Visible-light image (+IRCNN+VDSR) + syn-IR image (SI-AGAN) (proposed method)	12.95

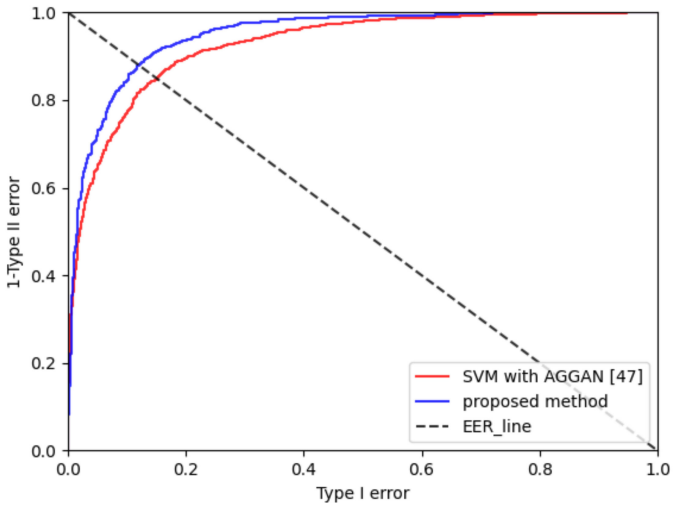


Figure 19. ROC curves of gender recognition accuracies using score-level fusion. (SVM with AGGAN [47] refers to Visible-light image(+IRCNN+VDSR) + syn-IR image (AGGAN) [47]).

Figure 20. Cases of Type I error, Type II error, correct recognition cases. (**a**) Cases of Type I error, (**b**) cases of Type II error, (**c**) correct recognition cases (male), (**d**) correct recognition cases (female). In (**a**–**d**), from left to right are original visible-light image, original IR image, reconstructed visible-light image, and syn-IR image.

Moreover, our proposed method was compared with previous methods where visible-light and IR images are combined. As shown in Table 16, our proposed method exhibited better performances than previous methods. Previous methods enhanced the gender recognition performance through handcrafted and CNN features. For a fair experiment, the study of [11] was divided into train, test, and validation, which are the same datasets as ours. In our proposed method, gender recognition performance was improved through SI-AGAN, which generated syn-IR images. Both visible-light and IR images are required for testing in previous methods, but gender recognition can be performed only with visible-light images in our proposed method.

Table 16. Comparison of gender recognition accuracies with our method and previous methods for the SYSU-MM01 database.

Methods	EER (%)
Visible-light image + IR image using HOG feature [4,22]	18.51
Visible-light image + IR image using weighted HOG feature [23]	23.90
Visible-light image + IR image using AlexNet [9,33]	24.53
Visible-light image (+IRCNN+VDSR) + IR image using ResNet-101 [11]	14.43
Visible-light image (+IRCNN+VDSR) + syn-IR image (SI-AGAN) using ResNet-101 (proposed method)	12.95

4.5. Computational Cost and Processing Time

Computational costs were measured and compared to prove that our proposed SI-AGAN reduced the computational cost than the original AGGAN. The average processing time was also measured and compared.

4.5.1. Computational Cost

The computational costs of AGGAN and SI-AGAN were compared using floating point operations (FLOPS) and parameters (Params). Two evaluation metrics, the total number of FLOPs and Params, were measured using the profile library provided by using the TensorFlow framework. The computational costs of SI-AGAN and the original AGGAN were compared. As explained above, the computation cost was reduced by changing the convolutional layer to the depthwise separable convolution layer and the perceptual loss was applied. For a quantitative comparison, Table 17 shows the comparison of FLOPS and the number of parameters between our proposed SI-AGAN and the original AGGAN. As shown in Table 17, our proposed SI-AGAN significantly reduced the computational cost compared to the conventional model. Accordingly, it was proven that our proposed SI-AGAN model has a lower computational cost and higher efficiency than other previous models.

Table 17. Comparison of FLOPs and parameters between AGGAN and SI-AGAN. #FLOPs and #Params refer to the total number of FLOPs and trainable parameters, respectively.

	#FLOPs	**#Params**
AGGAN [47]	9.19×10^9	2.95×10^6
SI-AGAN	2.28×10^9	5.40×10^5

4.5.2. Processing Time

The average processing time was measured and compared between our proposed SI-AGAN and the original AGGAN; the average processing time of our proposed method was also measured. The measurements were performed in a desktop environment and in the Jetson TX2 embedded system (NVIDIA Pascal ^{TM}-family GPU including 256 CUDA cores) [50]. Table 18 presents the average processing time of SI-AGAN and AGGAN in each environment. Compared to AGGAN, SI-AGAN had a shorter processing time by 1.72 ms on a desktop environment and by 20.56 ms on the Jetson TX2 environment. Our proposed SI-AGAN has a shorter processing speed than the original AGGAN

Table 18. Comparison of the average processing time between AGGAN and SI-AGAN. (unit: ms).

Environments	**AGGAN**	**SI-AGAN**
Desktop computer	18.01	16.29
Jetson TX2 embedded system	66.17	45.61

Table 19 presents the average processing time of our proposed method in desktop and Jetson TX2 environments. The average processing time is approximately 47.29 ms in a desktop environment and approximately 144.87 ms in the Jetson TX2 environment. The Jetson TX2 environment has a higher processing time than the desktop environment because the Jetson TX2 is an embedded system with limiting processing time.

Table 19. Average processing time of our proposed method (unit: ms).

Environments	**2-Step Image Reconstruction**	**SI-AGAN**	**ResNet-101 and Score-Level Fusion**	**Total**
Desktop computer	6.31	16.29	24.69	47.29
Jetson TX2 embedded system	8.48	45.64	90.78	144.90

5. Conclusions

We proposed a method for enhancing gender recognition in human images captured in uncontrolled environments. In most previous studies, gender recognition performance was limited because only visible-light images were used. Features are usually difficult to train due to excessive information such as background, accessories, clothes, and hair styles when training a recognizer. Also, there were many constraints to using both visible-light and IR images in previous research. NIR images require a separate NIR camera and NIR illuminator, and FIR camera is not widely used because of expensive equipment. Considering such facts, we proposed SI-AGAN that generated syn-IR images having similar characteristics as IR images. Because the syn-IR image generated by SI-AGAN has similar characteristics to the IR image, the performance degradation caused by various factors such as background, accessories, clothes, and hair styles was prevented. Our proposed SI-AGAN reduced computational costs by using a depthwise separable convolutional layer. This was proved by comparing the original AGGAN and our proposed SI-AGAN based on floating point operations and processing time. SI-AGAN not only reduced computational cost, but also showed higher performance than original AGGAN. Also, SI-AGAN used perceptual loss based on VGG Net-19 as well as pixel-based loss. Therefore, we improved the recognition performance of the generated image by considering the differences between the feature maps, and SI-AGAN generates relatively clear image compared to other various GAN models.

We combined the image generated through SI-AGAN with the visible-light image obtained through a two-step image reconstruction process to improve the gender recognition performance. By applying two-step image reconstruction, we improved the performance by reducing the influence of factors such as blur, noise, and low resolution, which degrade the performance of gender recognition.

In particular, our proposed method requires only visible-light images for conducting an experiment during the test step. We showed that our proposed method has superior performance to the state-of-the-art methods.

In future work, we will study different methods for improving quality even further by considering super resolution in addition to style transfer when generating images. Diverse pruning algorithms will be also applied to further reduce computational costs.

Author Contributions: Methodology, N.R.B.; Conceptualization, S.W.C.; Validation, J.H.K.; Supervision, K.R.P.; Writing—original draft, N.R.B.; 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-2020R1A2C1006179), 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: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Ng, C.-B.; Tay, Y.-H.; Goi, B.-M. A review of facial gender recognition. *Pattern Anal. Appl.* **2015**, *18*, 739–755. [CrossRef]
2. Yu, S.; Tan, T.; Huang, K.; Jia, K.; Wu, X. A study on gait-based gender classification. *IEEE Trans. Image Process.* **2009**, *18*, 1905–1910. [PubMed]
3. Patua, R.; Muchhal, T.; Basu, S. Gait-based person identification, gender classification, and age estimation: A review. *Prog. Adv. Comput. Intell. Eng.* **2021**, *1198*, 62–74.

4. Cao, L.; Dikmen, M.; Fu, Y.; Huang, T.S. Gender recognition from body. In Proceedings of the 16th ACM international Conference on Multimedia, Vancouver, BC, Canada, 26–31 October 2008; pp. 725–728.
5. Collins, M.; Zhang, J.; Miller, P.; Wang, H. Full Body Image Feature Representations for Gender Profiling. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, Kyoto, Japan, 27 September–4 October 2009; pp. 1235–1242.
6. Bourdev, L.; Maji, S.; Malik, J. Describing People: A Poselet-Based Approach to Attribute Classification. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1543–1550.
7. Guo, G.; Mu, G.; Fu, Y. Gender from Body: A Biologically-Inspired Approach with Manifold Learning. In Proceedings of the Asian Conference on Computer Vision, Xi'an, China, 23–27 September 2009; pp. 236–245.
8. Ng, C.-B.; Tay, Y.-H.; Goi, B.-M. A Convolutional Neural Network for Pedestrian Gender Recognition. In Proceedings of the International Symposium on Neural Networks, Dalian, China, 4–6 July 2013; pp. 558–564.
9. Antipov, G.; Berrani, S.-A.; Ruchaud, N.; Dugelay, J.-L. Learned vs. Hand-Crafted Features for Pedestrian Gender Recognition. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 1263–1266.
10. Ng, C.-B.; Tay, Y.-H.; Goi, B.-M. Pedestrian gender classification using combined global and local parts-based convolutional neural networks. *Pattern Anal. Appl.* **2019**, *22*, 1469–1480. [CrossRef]
11. Baek, N.R.; Cho, S.W.; Koo, J.H.; Truong, N.Q.; Park, K.R. Multimodal camera-based gender recognition using human-body image with two-step reconstruction network. *IEEE Access* **2019**, *7*, 104025–104044. [CrossRef]
12. Attention-Guided GAN for Synthesizing Infrared Image (SI-AGAN) and Syn-IR Datasets. Available online: http://dm.dgu.edu/link.html (accessed on 24 August 2021).
13. Althnian, A.; Aloboud, N.; Alkharashi, N.; Alduwaish, F.; Alrshoud, M.; Kurdi, H. Face gender recognition in the wild: An extensive performance comparison of deep-learned, hand-crafted, and fused features with deep and traditional models. *Appl. Sci.* **2021**, *11*, 89. [CrossRef]
14. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
15. Freund, Y.; Schapire, R.E. Experiments with a New Boosting Algorithm. In Proceedings of the 13th International Conference on Machine Learning, Bari, Italy, 3–6 July 1996; pp. 148–156.
16. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
17. Joachims, T. Making Large-Scale Support Vector Machine Learning Practical, Advances in Kernel Methods. Support Vector Learning. 1999. Available online: https://ci.nii.ac.jp/naid/10011961265/en/ (accessed on 1 October 2021).
18. Webb, A.R. *Statistical Pattern Recognition*; John Wiley & Sons: Hoboken, NJ, USA, 2003.
19. Cai, D.; He, X.; Han, J.; Zhang, H.-J. Orthogonal laplacianfaces for face recognition. *IEEE Trans. Image Process.* **2006**, *15*, 3608–3614.
20. Cai, D.; He, X.; Zhou, K.; Han, J.; Bao, H. Locality Sensitive Discriminant Analysis. In Proceedings of the International Joint Conference on Artificial Intelligence, Hyderabad, India, 6–12 January 2007; pp. 708–713.
21. Yan, S.; Xu, D.; Zhang, B.; Zhang, H.-J.; Yang, Q.; Lin, S. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *29*, 40–51. [CrossRef]
22. Nguyen, D.T.; Park, K.R. Body-based gender recognition using images from visible and thermal cameras. *Sensors* **2016**, *16*, 156. [CrossRef]
23. Nguyen, D.T.; Park, K.R. Enhanced gender recognition system using an improved histogram of oriented gradient (HOG) feature from quality assessment of visible light and thermal images of the human body. *Sensors* **2016**, *16*, 1134. [CrossRef]
24. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. In Proceedings of the 2012 Annual Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
25. 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.
26. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
27. Raza, M.; Sharif, M.; Yasmin, M.; Khan, M.A.; Saba, T.; Fernandes, S.L. Appearance based pedestrians' gender recognition by employing stacked auto encoders in deep learning. *Futur. Gener. Comp. Syst.* **2018**, *88*, 28–39. [CrossRef]
28. Cai, L.; Zhu, J.; Zeng, H.; Chen, J.; Cai, C.; Ma, K.-K. HOG-assisted deep feature learning for pedestrian gender recognition. *J. Frankl. Inst.* **2018**, *355*, 1991–2008. [CrossRef]
29. Fayyaz, M.; Yasmin, M.; Sharif, M.; Raza, M. J-LDFR: Joint low-level and deep neural network feature representations for pedestrian gender classification. *Neural Comput. Appl.* **2020**, *33*, 1–31. [CrossRef]
30. 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, 27–30 June 2016; pp. 770–778.
31. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person Re-Identification by Local Maximal Occurrence Representation and Metric Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2197–2206.
32. Wang, X.; Zheng, S.; Yang, R.; Zheng, A.; Chen, Z.; Tang, J.; Luo, B. Pedestrian attribute recognition: A survey. *Pattern Recognit.* **2021**, *121*, 108220. [CrossRef]

33. Nguyen, D.T.; Kim, K.W.; Hong, H.G.; Koo, J.H.; Kim, M.C.; Park, K.R. Gender recognition from human-body images using visible-light and thermal camera videos based on a convolutional neural network for image feature extraction. *Sensors* **2017**, *17*, 637. [CrossRef] [PubMed]
34. 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.
35. Cai, L.; Zeng, H.; Zhu, J.; Cao, J.; Wang, Y.; Ma, K.-K. Cascading Scene and Viewpoint Feature Learning for Pedestrian Gender Recognition. *IEEE Internet Things J.* **2020**, *8*, 3014–3026. [CrossRef]
36. Zhang, K.; Zuo, W.; Gu, S.; Zhang, L. Learning deep CNN Denoiser Prior for Image Restoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3929–3938.
37. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
38. Zhu, J.-Y.; Park, T.; 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. 2223–2232.
39. Hi-CMD. Available online: https://github.com/bismex/HiCMD (accessed on 24 August 2021).
40. Wu, A.; Zheng, W.S.; Yu, H.X.; Gong, S.; Lai, J. RGB-infrared Cross-Modality Person Re-Identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5380–5389.
41. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
42. Bottou, L. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 421–436.
43. Vedaldi, A.; Lenc, K. MatConvNet—Convolutional Neural Networks for MATLAB. In Proceedings of the ACM International Conference on Multimedia, Shanghai, China, 23–26 June 2015.
44. Tensorflow: The Python Deep Learning Library. Available online: https://www.tensorflow.org/ (accessed on 24 August 2021).
45. NVIDIA GeForce GTX 1070 Card. Available online: https://www.nvidia.com/en-in/geforce/products/10series/geforce-gtx-1070/ (accessed on 24 August 2021).
46. Kniaz, V.V.; Knyaz, V.A.; Hladuvka, J.; Kropatsch, W.G.; Mizginov, V. Thermalgan: Multimodal Color-to-Thermal Image Translation for Person Re-Identification in Multispectral Dataset. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
47. Mejjati, Y.A.; Richardt, C.; Tompkin, J.; Cosker, D.; Kim, K.I. Unsupervised attention-guided image to image translation. *arXiv* **2018**, arXiv:1806.02311.
48. Livingston, E.H. Who was student and why do we care so much about his t-test? *J. Surg. Res.* **2004**, *118*, 58–65. [CrossRef] [PubMed]
49. Cohen, J. A power primer. *Psychol. Bull.* **1992**, *112*, 155–159. [CrossRef] [PubMed]
50. Jetson TX2 Module. Available online: https://developer.nvidia.com/embedded/jetson-tx2 (accessed on 24 August 2021).