IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 20, 2020, accepted April 23, 2020, date of publication April 29, 2020, date of current version May 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2991255

A DCNN-Based Fast NIR Face Recognition System Robust to Reflected Light From Eyeglasses

JEYEON KIM[1], **MOONSOO RA**[2], **AND WHOI-YUL KIM**[1]
[1]Department of Electronics and Computer Engineering, Hanyang University, Seoul 04763, South Korea
[2]LightVision Inc., Seoul 04793, South Korea

Corresponding authors: Moonsoo Ra (ravicmoon@gmail.com) and Whoi-Yul Kim (wykim@hanyang.ac.kr)

This work was supported in part by the Samsung Electronics' University Research and Development Program, and in part by the Human Resources Program in Energy Technology Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry and Energy (MOTIE) of the Republic of Korea under Grant 20184030201970.

ABSTRACT Due to an increasing need for face recognition under poor lighting conditions, near infrared (NIR) face recognition based on deep convolutional neural networks (DCNN) has become an active area of research. However, in NIR face images of eyeglasses wearers, reflected light is generated around the eyes due to active NIR light sources, and it is one of the main contributors to performance degradation in NIR face recognition. In addition, there have to date been no attempts to lighten DCNN models for NIR face recognition. To solve these problems, we propose a DCNN-based fast NIR face recognition system which is robust to reflected light. This work has two main contributions: 1) We generated synthetic face images of individuals with and without eyeglasses using our proposed CycleGAN-based Glasses2Non-glasses (G2NG) data augmentation. We then constructed an augmented training database by adding the synthetic images, and the database helps to make the NIR face recognition system robust against reflected light. 2) A lightweight NIR FaceNet (LiNFNet) architecture was developed to reduce the computational complexity of the proposed system by adapting the depthwise separable convolutions and linear bottlenecks to VGGNet 16. The proposed architecture reduces the computation required, while improving the performance of NIR face recognition. Through the experiments reported in this paper, we verified that the proposed G2NG data augmentation improved the face recognition validation rate by 99.09% for NIR face images which have the reflected light from eyeglasses. Also, LiNFNet reduces the number of multiplication operations by 4.4×10^9 compared with VGGNet 16.

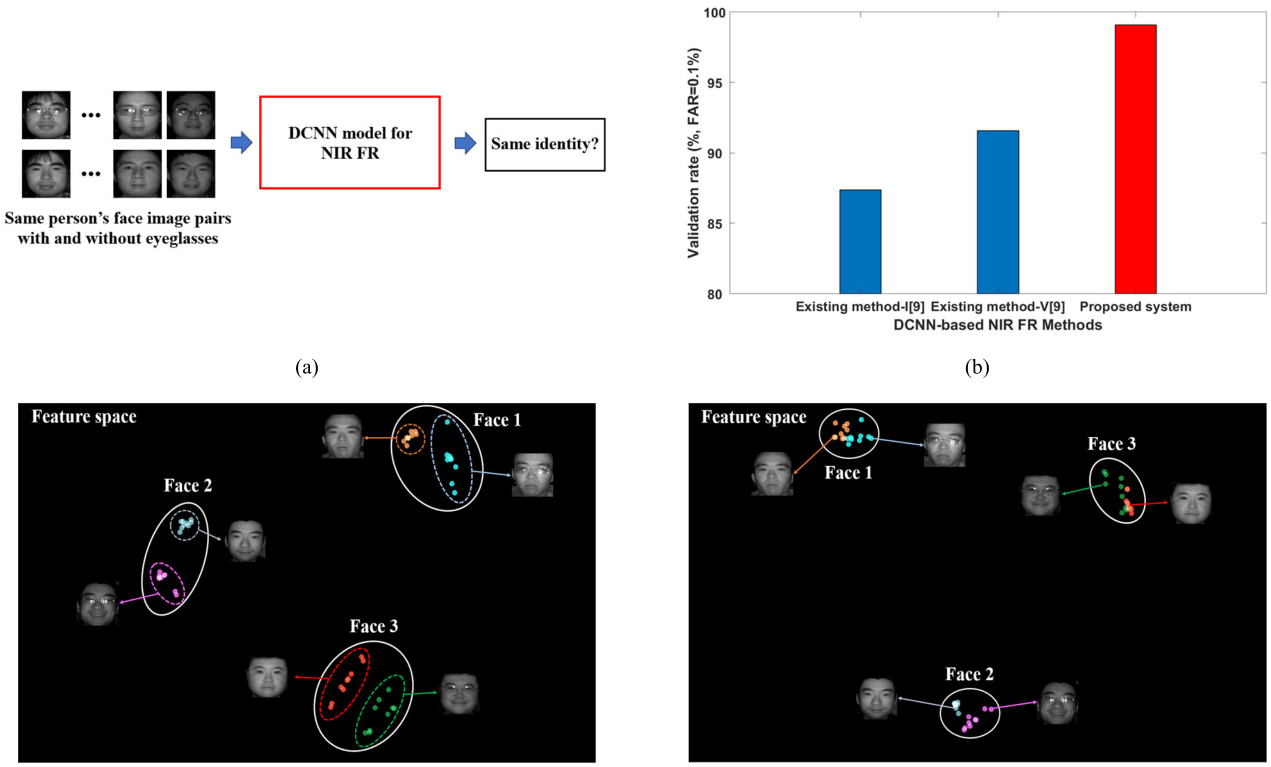
INDEX TERMS Biometrics, deep learning, NIR face identification, fine-tuning, lightweight deep CNN.

I. INTRODUCTION

Most deep convolutional neural networks (DCNN)-based face recognition (FR) studies have been conducted using RGB face images [1]-[8]. However, Kim *et al.* [9] showed that the validation rate of RGB FR decreases significantly under poor lighting conditions. In these environments, the validation rate of Kim's near infrared (NIR) FR method [9] was 40% or more higher than that of RGB FR. Since such environments are common in FR scenarios, such as unlocking a cell phone with FR in a dark room, it is important to research the field of NIR FR. Even though the Kim's method [9] has significantly improved the accuracy by introducing the fine-tuning approach into NIR FR, DCNN-based NIR FR still has considerable room for improvement with respect to accuracy and computational complexity.

One of the main issues with the existing NIR FR studies [9]–[11] is that their performances with respect to accuracy and validation rates are significantly reduced in Glasses and Non-glasses (G-NG) positive NIR FR scenarios. As shown in Fig. 1 (a), the scenario means that the system conducts the NIR FR for the face image pair of a person with and without eyeglasses. In this scenario, the validation rate is decreased because the gallery and probe images have large intensity differences around the eye regions due to reflected light. The validation rates of the Kim's method [9] are less than 93% in the scenario, as shown in Fig. 1 (b). Performance at this level cannot guarantee sufficient security to justify the use of NIR FR in the real world. Since G-NG positive NIR FR scenarios are very common in real-world applications, improving the performance of FR in such scenarios is crucial.

The associate editor coordinating the review of this manuscript and approving it for publication was Weizhi Meng.



(a) (b)

(c) (d)

FIGURE 1. (a) The G-NG positive NIR FR scenarios. (b) The validation rates of the Kim's method [9] and proposed method in the G-NG positive NIR FR scenarios. "Existing method-I" and "Existing method-V" are the Inception ResNet v1 and VGGNet 16 versions of the Kim's method [9], respectively. (c) and (d) show deep features of the Kim's method [9] and the proposed method for same person's face images with and without eyeglasses. These deep features are represented using t-SNE [12].

Another issue with the existing approaches is computational complexity. Despite recent advances in NIR FR [9]–[11], there are very few studies related to reducing the computational costs of NIR FR. Since recently-produced smartphones provide a feature that enables the unlocking of a phone using a face, it would be beneficial to make a lightweight and fast DCNN architecture for NIR FR.

In consideration of the above-mentioned issues, our goal was to develop a fast DCNN-based NIR FR system robust to reflected light. To achieve this objective, we utilized two contributions to construct the proposed NIR FR system:

1) CycleGAN-based Glasses2Non-glasses (G2NG) data augmentation
2) Lightweight NIR FaceNet (LiNFNet) architecture

The first contribution makes the DCNN architecture for NIR FR be trained robust against reflected light. The second contribution not only effectively reduces the computational cost of NIR FR, but also models human faces well even if reflected light is present. The detail explanations of the contributions are as follows.

A. CycleGAN-BASED G2NG DATA AUGMENTATION

When using publicly available NIR face databases to train DCNN architectures, we cannot adequately cover G-NG positive FR scenarios. This is because the numbers of face images both with and without eyeglasses are not balanced in most face labels of the public NIR face training databases. To solve an unbalanced data problem, three methods are frequently used: under-sampling [13]–[15], over-sampling [15], and synthetic over-sampling [16], [17]. If synthetic over-sampling methods can generate images close to real ones, we can increase a proportion of minorities in the database better than other sampling methods. In this point of view, we adapted CycleGAN to implement synthetic over-sampling, and generated realistic face images of individuals with and without eyeglasses.

B. LiNFNet ARCHITECTURE

Recently, several architectures [18]–[24] have been developed to reduce the computational cost of problems such as classification and detection, while maintaining accuracy. However, it is not clear that such architectures can achieve state-of-the-art performance in NIR FR. Instead of using the successful architectures [18]–[24] in classification or detection, we aimed to improve VGGNet 16 [25] and Inception ResNet v1 [26] known to have good performances in NIR FR. By adapting the depthwise separable convolutions [18] and linear bottlenecks [21] that efficiently reduce the number of

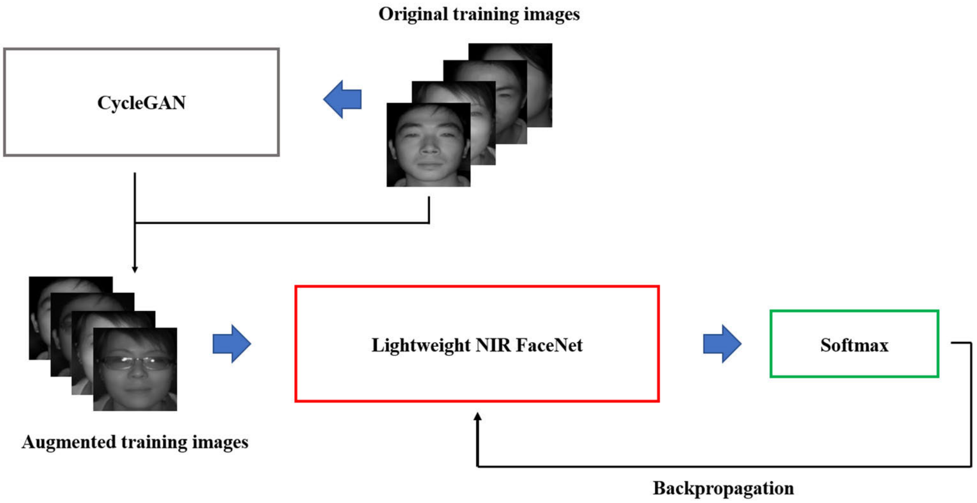


FIGURE 2. The training process of the proposed NIR FR system.

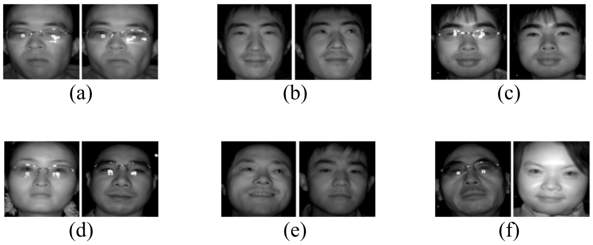


FIGURE 3. The types of input pairs according to the combination of face images with and without eyeglasses. (a) Glasses positive pair. (b) Non-glasses positive pair. (c) Mixed positive pair. (d) Glasses negative pair. (e) Non-glasses negative pair. (f) Mixed negative pair.

parameters and computations of convolution filters, we created a lightweight architecture for NIR FR, and we call this architecture LiNFNet in this paper.

To visualize the effect of two contributions on reflected light, we investigated the deep features used for NIR FR in the feature space using t-SNE [12]. The deep features produced by the proposed method, when applied to images of the same person wearing or not wearing eyeglasses, have less variance than those produced by Kim's method [9] as shown in Fig 1 (c) and (d). The discriminative ability of Kim's method [9] is acceptable for the three identities (Fig 1 (c)). However, NIR FR was conducted on the database, which includes more than two hundred identities, and the feature space is densely filled with the features from the identities. In this case, even slight distances between the features of face images of the same person with and without eyeglasses are likely to reduce the performance of NIR FR. In other words, the same identity's concentrated features produced by the proposed method contribute to improve the NIR FR performance in the G-NG positive FR scenario, and it can be found in Fig. 1 (b).

The rest parts of this paper are organized as follows. In Section II, related works of the proposed system are explained. Section III elaborates training and inference processes of the proposed system. CycleGAN-based G2NG data augmentation and LiNFNet are described in Section IV and V, respectively. In Section VI, the experimental results are presented. In Section VII, we conclude our work by summarizing the pros and cons of the proposed NIR FR system, and discussing the future works.

II. RELATED WORK

In this section, we summarize work related to the proposed NIR FR system's two contributions, the CycleGAN-based data augmentation and LiNFNet.

A. GAN-BASED DATA AUGMENTATION

Following the pioneering work of LeCun *et al.* [27] and Krizhevsky *et al.* [28], DCNN [25], [26], [29]–[31] became a main-stream approach to research into well-known computer vision problems such as recognition, classification, and segmentation. Using powerful deep models [25]–[31], performance on these problems has been drastically improved. However, such deep networks require numerous well-annotated databases to achieve state-of-the-art performance. Since obtaining such high-quality databases is time-consuming and expensive, data augmentation methods which generate synthetic training images have been actively researched. Recently, several studies [32]–[35] have utilized GAN [36]–[41] for data augmentation, and have succeeded in generating realistic synthetic training images.

DA-GAN [32] introduced the GAN architecture for instance-level image translation. In one example, synthetic bird images involving various poses were generated, and these images were used as training data for fine-grained classification.

Antoniou *et al.* [33] introduced a conditional GAN for data augmentation. From the encoder of the conditional GAN, a representation of the input image was acquired. The representation and a random vector were then concatenated, and the decoder generated a synthetic image

TABLE 1. The details of the G-NG mixed face classes and NIR face images in the public NIR face databases.

Public NIR face databases	# of G-NG mixed face classes	The ratio of G-NG mixed face classes to the total classes (%)	# of NIR face images
CASIA NIR [46]	64	32.5	3,938
CASIA VIS-NIR 2.0 [48]	86	11.9	12,485
PolyU-NIRFD [47]	2	0.9	24,698

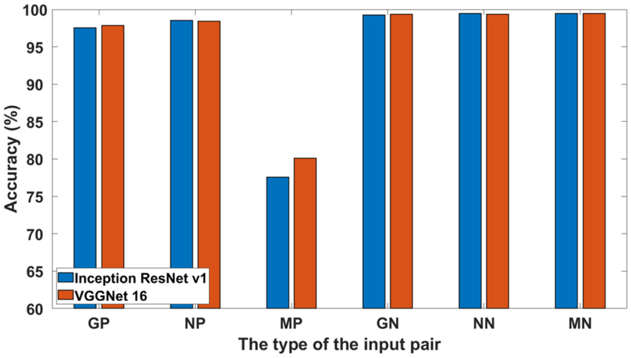


FIGURE 4. The accuracies of NIR FR for the input pair types (using CASIA VIS-NIR 2.0 [48] database for training). GP, NP, and MP are abbreviations for glasses, non-glasses, and mixed positive pairs, respectively. GN, NN, and MN also mean glasses, non-glasses, and mixed negative pairs.

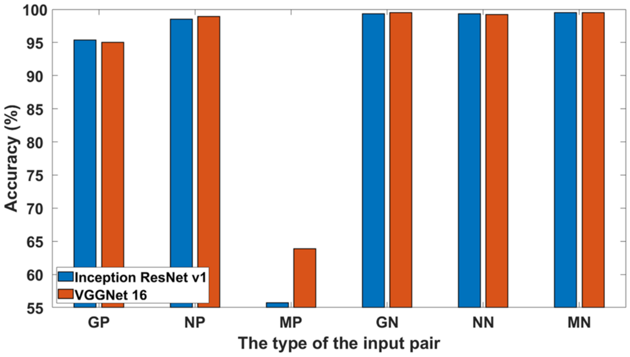


FIGURE 5. The accuracies of NIR FR for the input pair types (using PolyU-NIRFD [47] database for training).

from the concatenated vector. Using the conditional GAN, Antoniou *et al*. [33] constructed augmented databases for the Omniglot [42], EMNIST [43], and VGG-Face [1] databases. Antoniou *et al*. [33] showed that recognition accuracy was improved on these databases.

AugGAN [34] added a segmentation network to GAN to maintain the structures of the input images in the synthetic images.

FaceID-GAN [35] introduced the concept of three players: a generator, a classifier for identity classification, and a discriminator. With the training of the three players, the classifier for identity classification achieved high performance. Due to the classifier, the generator generated synthetic images while preserving the identities of the faces in the input images. Using the Shen's method [35], synthetic frontal face images were generated from face images which had various poses, and face verification was conducted using the synthetic images. Shen *et al.* [35] improved the verification accuracy.

To prevent degradation of the NIR FR performance due to reflected light, Jo and Kim [58] added the simple reflected light patterns, such as rectangle, circle, or ellipse shapes, to the parts of the NIR face images near the eyes. Although their data augmentation method improved the NIR FR performance, this approach did not generate the sufficiently realistic reflected light patterns in the NIR face images.

After reviewing the existing methods [32]–[35], [58], we postulated that there could be a performance improvement in NIR FR in G-NG positive FR scenarios when G2NG data augmentation was well conducted using GAN. In this work, since G2NG data augmentation can be represented as an unpaired image-to-image translation problem, we utilized CycleGAN [44] to generate synthetic images. We demonstrated that the NIR FR accuracy in the G-NG positive FR scenarios was improved using CycleGAN-based G2NG data augmentation, as shown in Section VI.

B. LIGHTWEIGHT DCNN MODELS

Despite the high accuracy of most DCNN-based applications, they cannot be applied in most smartphones or embedded environments, due to limited computing resources. To extend deep learning applications to mobile environments, it is necessary to conduct studies into the reduction of computational cost, by making the DCNN models lightweight. Also, there have been several studies [18], [19], [21]–[23] addressing this problem.

MobileNet v1 [18] introduced depthwise separable convolution to lighten the DCNN architecture. In the work of Howard *et al.* [18], ImageNet classification accuracy did not decrease significantly, while the computational burden was considerably reduced. Chollet [19] demonstrated that depthwise separable convolutions could be adapted to the inception modules [30]. The training speed of the Chollet's lightweight architecture [19] was increased compared to Inception v3 [30]. ShuffleNet v1 [23] utilized pointwise group convolutions to reduce the computational cost of pointwise convolutions and developed channel shuffle to overcome the side effect of pointwise group convolutions. Channel shuffle made it possible to transfer information between groups of activation channels. MobileNet v2 [21] developed a DCNN architecture with linear bottlenecks. Linear bottlenecks helped the efficient reduction of the channels

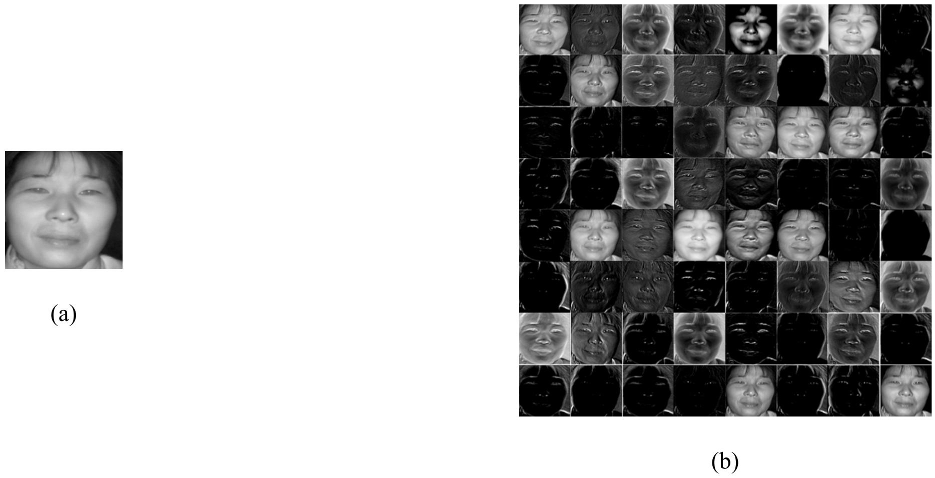


FIGURE 6. The output activations extracted from the first convolution layers of VGGNet 16 [25] for a NIR face image. (a) An input NIR face image. (b) shows the output activations of the first convolution layer. We normalized the intensity values of the output activations to be between 0 and 255.

TABLE 2. The NIR FR performance of Inception_Resnet_v1 [26] and VGGNet 16 [25]. CASIA VIS-NIR 2.0 [48] is utilized as the training database for fine-tuning, and the validation database is same as the test pairs in Fig. 4 and 5. The NIR FR was conducted on NVIDIA GTX 1080ti GPU. "Time" means the average time which is taken to extract features for NIR FR

	Inception Resnet v1	VGGNet 16
Accuracy (%)	97.78	98.53
Time (ms)	13.1	5.32

TABLE 3. The NIR FR accuracy of VGGNet 16 and VGGNet 16_light. VGGNet 16_light is the lightweight version of VGGNet 16. The number of filters in the first convolution layer of VGGNet 16_light is half of that of VGGNet 16. We constructed training databases for VGGNet 16 and VGGNet 16_light by integrating CASIA VIS-NIR 2.0 [48] and PolyU-NIRFD [47].

	VGGNet 16	VGGNet 16_light
Accuracy (%)	97.9	97.47

of the output activation, by estimating the manifold of the activation while retaining the information in the activation. In ShuffleNet v2 [22], channel split was introduced into the architecture that was introduced in ShuffleNet v1, to efficiently use the architecture.

Wu *et al.* [56] developed a light DCNN architecture for FR. They introduced max-feature-map (MFM) into each convolution layer, which helped their DCNN architecture to extract a compact face representation while reducing the number of parameters, and the computational costs. However, Wu's architecture [56] was not designed for NIR FR, and Wu *et al.* [56] did not sufficiently analyze the effects of reflected light in NIR face images on the performance of NIR FR. Zheng and Zu [57] developed a light DCNN architecture for RGB FR by adding a normalized layer to Wu's architecture [56]. Zheng's architecture [57], therefore, was also not designed for NIR FR.

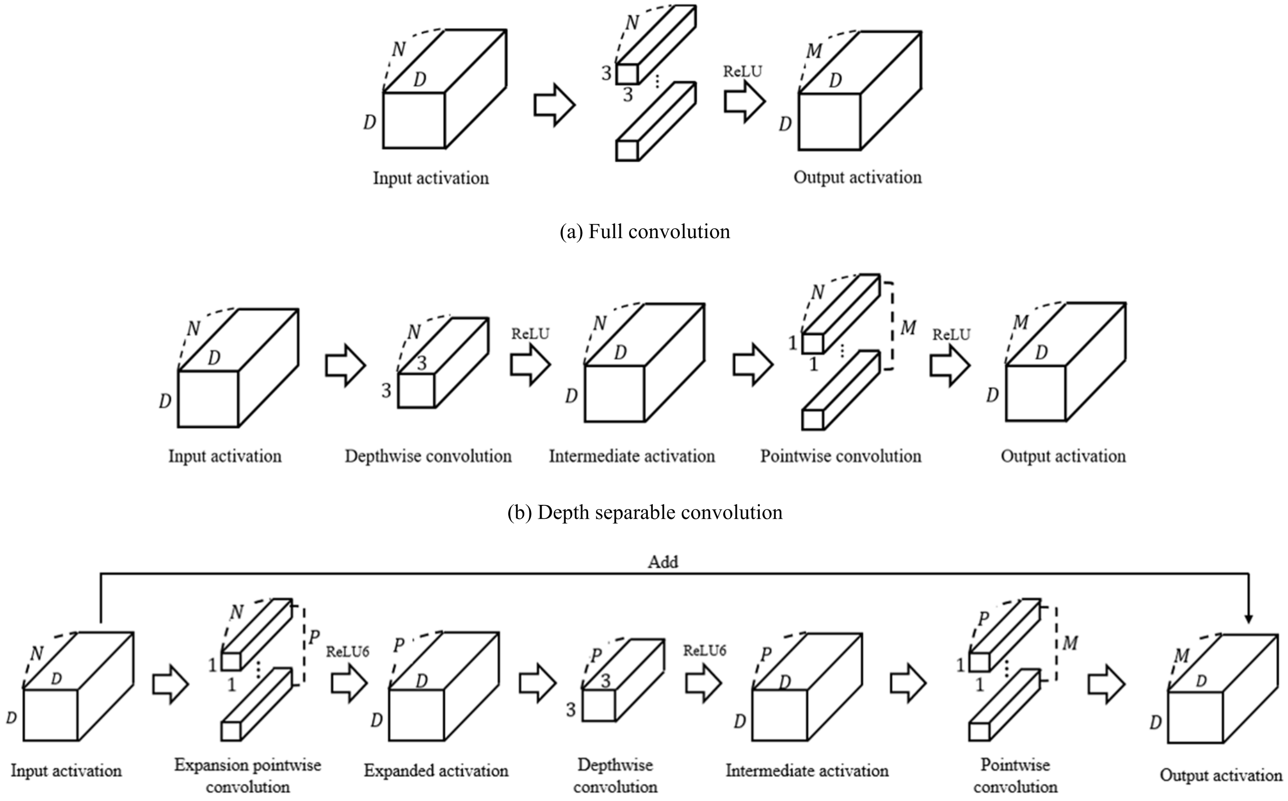
Since the need to use FR in smartphones and embedded environments is increasing, DCNN models for NIR FR should be lightened, as has previously been demonstrated for existing methods [18], [19], [21]–[23], [56], [57].

In the work reported in this paper, we lightened one of the powerful off-the-shelf DCNN architectures, VGGNet 16 [25]; this architecture was shown to have high performance for NIR FR in the literature [9]. The reason for using VGGNet 16 as a backbone network is that, in our toy experiment, VGGNet 16 is about twice as fast as another powerful architecture, Inception ResNet v1 [26]. In addition, the NIR FR accuracy of VGGNet 16 in G-NG positive FR scenarios is higher than that of Inception ResNet v1. The results of the toy experiment can be found in Section V. We lightened VGGNet 16 by simultaneously adapting depthwise separable convolutions [18] and linear bottlenecks [21]; the proposed lightweight model is called LiNFNet. Depthwise separable convolutions and linear bottlenecks significantly reduced the computational complexity of VGGNet 16. Especially, linear bottlenecks considerably improved the accuracy of NIR FR by efficiently increasing the number of channels of the input activations using pointwise convolutions.

III. PROPOSED NIR FR SYSTEM

An overview of the proposed system is presented in this section. The proposed system was designed as an end-to-end framework which includes the LiNFNet architecture. The inference process of the proposed system is same as FaceNet [2]:

1) A face image pair is inserted to our NIR FR system, and two deep features are extracted from the LiNFNet architecture.
2) Euclidean distance between the two features is calculated.



(a) Full convolution

(b) Depth separable convolution

(c) Linear bottleneck

FIGURE 7. The flow chart of the three convolution modules. *D* is the dimension of the activations. *N*, *P*, and *M* are the number of the channels of the input, expanded, and output activation, respectively. In the linear bottlenecks of LiNFNet, "Add" operator is conducted for the residual connection. When the number of channels of the input and output activations is different, a full convolution is used for residual connection.

3) If the distance is less than a predefined threshold, the system considers that the two face images are from the same identity; otherwise, the images are from different identities.

In Fig. 2, the training process of the proposed NIR FR system is depicted. Before training the LiNFNet architecture, G2NG data augmentation is conducted to robustly train LiNFNet against reflected light from eyeglasses. During the data augmentation, CycleGAN [44] generates synthetic NIR face images of individuals with and without eyeglasses. Then, we construct the augmented training database by merging the real and synthetic images. The numbers of the face images with and without eyeglasses in the augmented database are balanced. According to Kim *et al.* [9], the fine-tuning approach to NIR FR achieved a better validation rate than the learning from scratch approach. As with the fine-tuning approach of Kim *et al.* [9], we utilized a pretrained model of LiNFNet on CASIA WebFace [45] and, conducted fine-tuning on the augmented training database.

IV. CYCLEGAN-BASED G2NG DATA AUGMENTATION

A. MOTIVATION

After reviewing publicly available NIR face images, we predicted that the accuracy of NIR FR would be decreased in the G-NG positive FR scenarios due to reflected light. To investigate this hypothesis, we defined six types of input pairs as shown in Fig. 3, and conducted two toy experiments.

In Fig. 3, the input pairs containing 0, 1, and 2 eyeglasses wearers are denoted as ''non-glasses'', ''mixed'', and ''glasses'', respectively. If the input pair was taken from one person, we denoted it as a ''positive'' pair; otherwise, it is a ''negative'' pair. Therefore, mixed positive pairs are identical to the G-NG positive FR scenarios.

Through the two toy experiments, we evaluated the NIR FR accuracies of the six types of input pairs. For each type of input pair, we extracted 2,000 pairs from the CASIA NIR [46] database, producing a total of 12,000 pairs for evaluation. The first and second experiments used CASIA VIS-NIR 2.0 [48] and PolyU-NIRFD [47], respectively, as training databases for the fine-tuning approach. In both experiments, we utilized Inception ResNet v1 and VGGNet 16 as backbone networks for the NIR FR system. The results of the experiments are summarized in Fig. 4 and Fig. 5.

As shown in Fig. 4, all types of input pairs except for the mixed positive pairs achieved an accuracy of more than 97%. On the other hand, the mixed positive pair achieved an accuracy of about 80%. This phenomenon can also be seen in Fig. 5. From these observations, we can say that the G-NG positive FR scenarios caused a number of failure cases in NIR FR due to the reflected lights from eyeglasses.

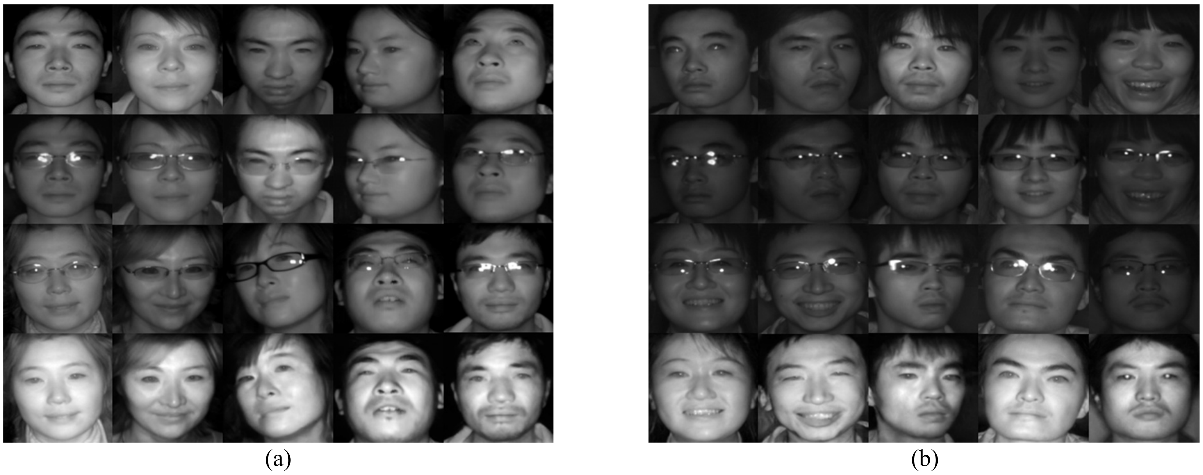


FIGURE 8. The success cases of the proposed CycleGAN-based G2NG data augmentation. The first and third rows show real NIR face images without and with eyeglasses, respectively. In the second and fourth rows, the synthetic NIR face images with and without eyeglasses are shown, respectively. (a) The result images of CASIA VIS-NIR 2.0 [48] database. (b) The result images of PolyU-NIRFD [47] database.

TABLE 4. The architecture of LiNFNet. *d* and *c* is the number of the channels in the expanded and output activation in Fig. 7. *r* and *s* are the number of the repeated times and strides.

Input	Operator		*d*	*c*	*r*	*s*
160^2 x 3	full conv2d		-	32	1	1
160^2 x 32	depthwise separable		-	32	1	1
160^2 x 32	maxpool 2x2		-	32	1	2
80^2 x 32	depthwise separable		-	128	2	1
80^2 x 128	maxpool 2x2		-	128	1	2
40^2 x 128	linear bottleneck		512	256	3	1
40^2 x 256	maxpool 2x2		-	256	1	2
20^2 x 256	linear bottleneck		768	512	3	1
20^2 x 512	maxpool 2x2		-	512	1	2
10^2 x 512	linear bottleneck		768	512	3	1
10^2 x 512	maxpool 2x2		-	512	1	2
5^2 x 512	avgpool 5x5		-	512	1	1
1x512	Dropout		-	512	1	-
1x512	FC		-	128	1	-

TABLE 5. The number of the training data for the CycleGAN-based G2NG data augmentation. "Glasses" and "Non-glasses" means the NIR face images with and without eyeglasses for training, respectively.

Training database	Glasses	Non-glasses
CASIA VIS-NIR 2.0 [48]	1,011	4,994
PolyU-NIRFD [47]	6,062	6,042
Total	7,073	11,036

To reduce the number of failure cases, each face label in the training NIR face databases should include a number of face image pairs with and without eyeglasses, and the number of these two types of face images should be similar. In other words, the databases should have a number of Glasses and Non-glasses (G-NG) mixed face classes; the G-NG mixed face classes denotes face classes that contain both face images with and without eyeglasses. In Table 1, information about G-NG mixed face classes and total face images in several public NIR databases [46][48] is presented. The CASIA VIS-NIR 2.0 [48] database has 86 G-NG mixed face classes. However, in this database, the ratio of G-NG mixed face classes to all face classes is low, at 11.9%. The PolyU-NIRFD [47] database has only two G-NG mixed face classes. Therefore, we expect that a DCNN model trained using the PolyU-NIRFD [47] and CASIA VIS-NIR 2.0 [48] databases will not be robust to G-NG positive FR scenarios. As shown in Table 1, the ratio of G-NG mixed face classes to all face classes is 32.5% in the CASIA NIR database [46]. Although this ratio is the highest among the databases summarized in Table 1, the CASIA NIR database is unsuitable for training DCNN models for NIR FR because there are only about 4,000 face images in the database. Therefore, G2NG data augmentation should be carried out to increase the number of G-NG mixed face classes in the CASIA VIS-NIR 2.0 and PolyU-NIFRD databases.

B. CYCLEGAN FOR G2NG DATA AUGMENTATION

The objective of the G2NG data augmentation is to produce both synthetic face images with and without eyeglasses. To make synthetic face images with eyeglasses, reflected light should be added; otherwise, reflected light should be removed. This objective can be achieved by solving an image-to-image translation problem.

As compared to the well-known Pix2Pix [49] which solves the paired image-to-image translation problem, CycleGAN [44] has two advantages. Firstly, it does not require paired annotations; it only requires images from two domains. Secondly, it can learn to produce outputs from both domains (A2B and B2A). These two advantages are crucial for our

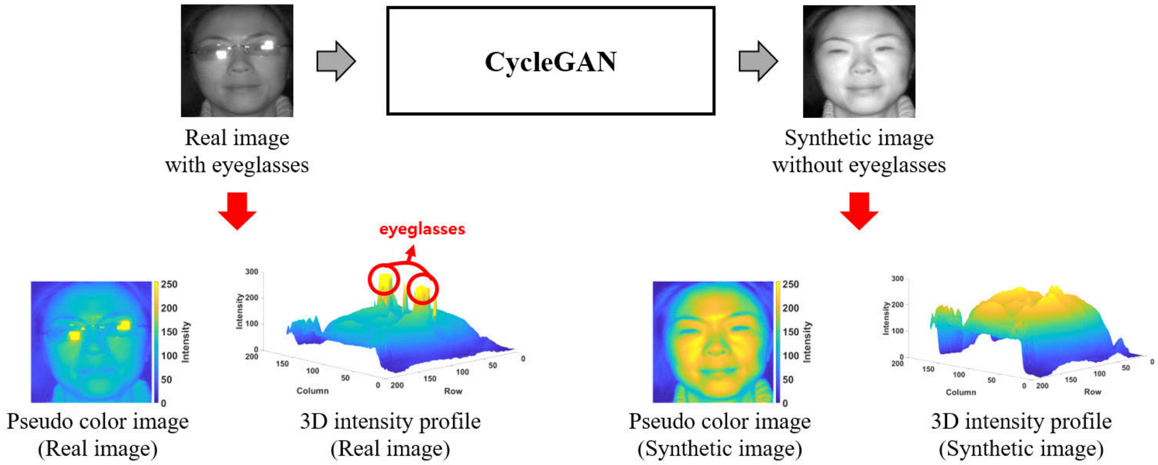


FIGURE 9. The pseudo color images and 3D intensity profiles of the real image with eyeglasses and the synthetic image without eyeglasses.

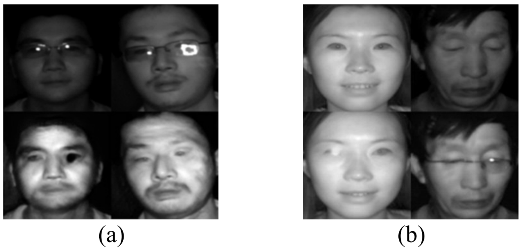


FIGURE 10. The failure cases of the proposed CycleGAN-based G2NG data augmentation. The first row shows the real NIR face images, and the second row shows the synthetic NIR face images. (a) The failed examples of the synthetic NIR face images without eyeglasses (b) The failed examples of the synthetic NIR face images with eyeglasses.

TABLE 6. The number of the NIR face images in the training databases for the quantitative evaluation of the CycleGAN-based G2NG data augmentation. "AUG" means that the corresponding database is the augmented database which is constructed by the proposed data augmentation.

Training database	# of NIR face images
CASIA VIS-NIR 2.0 [48]	12,485
PolyU-NIRFD [47]	24,698
CASIA VIS-NIR 2.0_AUG	24,970
PolyU-NIRFD_AUG	49,396

application, because it is very difficult to acquire paired NIR face images with and without eyeglasses. Therefore, we used CycleGAN [44] rather than Pix2Pix [49] for the G2NG data augmentation.

To train CycleGAN for G2NG data augmentation, we used the same architecture and loss as in Zhu *et al.* [44], and identity loss [44] was also utilized to preserve the identities while generating the synthetic face images with and without eyeglasses. The images resulting from the CycleGAN-based G2NG data augmentation are shown in Fig. 8 and Fig. 10 in Section VI.

V. LINFNET ARCHITECTURE

In Kim *et al.* [9], it was shown that Inception ResNet v1 [26] and VGGNet 16 [25] achieved a high validation rate for NIR FR. Therefore, we expected that making lightweight versions of Inception ResNet v1 or VGGNet 16 would be effective. As shown in Table 2, VGGNet 16 is about 2.5 times faster than Inception ResNet v1; hence, VGGNet 16 is a more suitable architecture than Inception ResNet v1 for the proposed NIR FR system. In addition, VGGNet 16 has an advantage that its accuracy is higher than that of Inception ResNet v1 in the G-NG positive FR scenarios. Because of the NIR FR accuracy and speed, we chose VGGNet 16 to make a lightweight DCNN architecture for NIR FR.

In this study, we produced LiNFNet by lightening VGGNet 16 [25] using depthwise separable convolutions [18] and linear bottlenecks [21]. When constructing the LiNFNet architecture, we decreased the number of filters in the first convolution layer of the network by half. Fig. 6 shows several output activations extracted from the first convolution layers of VGGNet 16 for an NIR face image. These activations have similar patterns and structures of the intensity values. From this observation, we conclude that the activations contain redundant information. Thus, decreasing the number of convolution filters in the first layer does not significantly decrease the NIR FR accuracy. The result of such reduction is shown in Table 3.

We made the initial convolution layers of LiNFNet by adapting the depthwise separable convolutions [18] to the 2nd, 3rd, and 4th convolutions of VGGNet 16. We expected that the NIR FR accuracy would not significantly decrease upon replacing the full convolutions of the initial convolution layers with the depthwise separable convolutions [18], which are the lightweight version of full convolutions. This is because the initial convolution layers are simpler functions for extracting the output activations than the rest of the

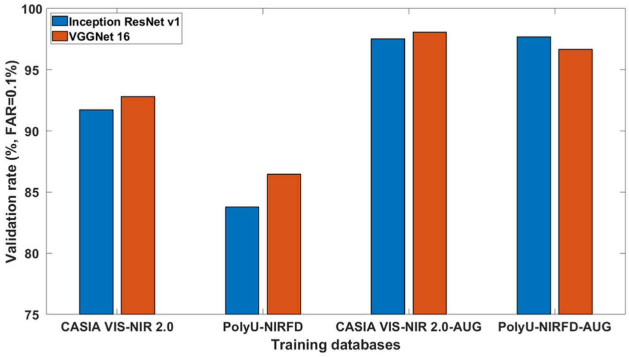


FIGURE 11. The validation rate of NIR FR on the original and augmented training databases.

convolution layers; the initial convolutions extract the low-level information, such as the edges and the combination of the edges, for the input NIR face images. From the experiment reported in this paper, we found that such replacement effectively reduces the computational complexity while improving the NIR FR accuracy in the G-NG positive FR scenarios.

It is necessary that the layers following the initial convolution layers extract rich feature information for NIR FR from the input activation. To produce output activations including this rich information, we should expand the input activation by increasing the number of channels, and extract the output activation by combining many channels of the expanded input activation. However, as the number of channels of the input activation increases, the computational complexity also increases. Therefore, we should efficiently extract the rich information for NIR FR from the input activation while preserving a low computational complexity. To do this, we adapted linear bottlenecks [21] to the last three convolution layers of VGGNet 16 to make the LiNFNet architecture.

In Fig. 7 (c), the expansion pointwise convolution of the linear bottleneck increases the number of channels of an input activation to extract the rich information for NIR FR. The depthwise convolution of the linear bottleneck extracts the rich information for each channel of the input activation. Pointwise convolution linearly decreases the number of channels of the output activation to reduce the computational cost of the next convolution layer. This approach helped us to efficiently extract more rich information for NIR FR than full convolution or depthwise separable convolution. As explained in Sandler *et al.* [21], the information in the intermediate activation in Fig. 7 (c) is considerably redundant for NIR FR. Therefore, the number of channels of intermediate activation can be linearly reduced using pointwise convolution. To prevent information loss, we did not use ReLU6 after the pointwise convolution in the same manner as Sandler *et al.* [21]. Since the manifold of the output activation can be well acquired by linearly reducing the number of channels of output activation, additional information loss from ReLU6, which is a nonlinear function, causes a considerable drop in the NIR FR accuracy. The LiNFNet architecture is summarized in Table 4.

TABLE 7. The accuracy of NIR FR for the mixed positive pairs in Fig. 3 according to the original and augmented training databases.

	Inception ResNet v1	VGGNet 16
CASIA VIS-NIR 2.0 [48]	77.55	80.1
CASIA VIS-NIR 2.0_AUG	94.8	96.35
PolyU-NIRFD [47]	55.75	63.9
PolyU-NIRFD_AUG	95.9	93.9

It is necessary to compare the computational complexity of a full convolution, depthwise separable convolution [18], and linear bottleneck [21] to verify the extent to which LiNFNet reduces the number of computations compared with VGGNet 16 [25]. In this paper, only the multiply operation is considered. The equations to compute the number of multiply operations in the convolution modules are as follows:

$$C_F = 9ND^2M, \tag{1}$$
$$C_D = 9ND^2 + ND^2M, \tag{2}$$
$$C_L = ND^2P + 9D^2P + PD^2M, \tag{3}$$

where these equations can be derived from Fig. 7. C_F, C_D, and C_L are the numbers of multiply operations of a full convolution, depthwise separable convolution, and linear bottleneck, respectively. The meanings of the other notations are shown in Fig. 7. Equations (1) and (2) were formulated in the literature [18].

To quantitatively verify how much lighter LiNFNet is than VGGNet 16 [25], we calculated the differences (D_D) between the number of the multiply operations of the full convolution and depthwise separable convolution. For the linear bottleneck, we calculated D_L in the same manner as the linear bottleneck.

$$D_D = C_D - C_F = ND^2(9 - 8M), \tag{4}$$
$$D_L = C_L - C_F = D^2(NP + 9P + PM - 9NM), \tag{5}$$

If D_D or D_L have negative values, the number of multiply operations of the depthwise separable convolution or linear bottleneck will be lower than that of the full convolution, and vice versa. From equations (4) and (5), the number of the multiply operations of LiNFNet is about 4.4×10^9 lower than that of VGGNet 16.

VI. EXPERIMENTS

In this section, we evaluated performance of LiNFNet regarding robustness against reflected light and performance versus computational complexity trade-off. In addition, competitive analysis of the proposed system with existing systems [9]–[11] was conducted. For the two main experiments, the augmented database, which was constructed by CycleGAN-based G2NG data augmentation, should be utilized. Therefore, before the main experiments, we conducted the qualitative and quantitative evaluations of the proposed data augmentation.

TABLE 8. The database configuration according to the experiments.

The title of the experiment	Training database	Validation or test database
Performance evaluation of LiNFNet	- Integrated NIR Face database • CASIA VIS-NIR 2.0 + PolyU • 948 identities • 37,183 NIR face images - Integrated NIR Face_AUG database • The augmented version of Integrated NIR Face database • 948 identities • 70,175 NIR face images - These databases are used to show the effect of the CycleGAN-based G2NG data augmentation	- CASIA NIR face database is used as validation database - We extract 12,000 validation pairs from the validation database (Verification scenario) • 2000 glasses positive pairs • 2000 non-glasses positive pairs • 2000 mixed positive pairs • 2000 glasses negative pairs • 2000 non-glasses negative pairs • 2000 mixed negative pairs
Performance comparison of the proposed NIR FR system and existing methods	Same as above	- Glasses2Non-glasses (G2NG) test database • The NIR FR with this database is identical the G-NG positive identification scenario • Gallery images: 192 face images with eyeglasses (64 identities x 3 images) • Probe images: 320 face images without eyeglasses (64 identities x 5 images)

In Section IV-A, the qualitative and quantitative evaluations of the proposed data augmentation are described. Databases and training setup for the two main experiments are present from Section IV-B and IV-C, respectively. In Section IV-D and IV-E, the descriptions of the main experiments are provided.

A. CYCLEGAN-BASED G2NG DATA AUGMENTATION

In these experiments, qualitative and quantitative performance evaluations were conducted for the CycleGAN-based G2NG data augmentation.

1) QUALITATIVE EVALUATION

Through the performance evaluation, we investigated how realistically the proposed G2NG data augmentation generates the synthetic NIR face images with and without eyeglasses from real images. We split the CASIA VIS-NIR 2.0 [48] and PolyU-NIRFD [47] databases into training and test databases. Table 5 shows the number of training face images with and without eyeglasses. For testing, we used all of the NIR face images in the CASIA VIS-NIR 2.0 and PolyU-NIRFD databases. In Fig. 8 and 10, the results of the proposed CycleGAN-based G2NG data augmentation are shown.

In Fig. 8, the synthetic images with and without eyeglasses are very similar to real images. In the synthetic images with eyeglasses, the reflected lights, which are generated around the eyes, appear in various patterns. Therefore, the generalization ability of CycleGAN is good with respect to the generation of various reflected lights. Even though the average intensity values of the synthetic images without eyeglasses were higher than those of the real images, the reflected lights of the real images with eyeglasses were successfully removed in the synthetic images, and the identities of the real images are well preserved in the synthetic images. To analyze the phenomenon in which synthetic images without eyeglasses are brighter than real images with eyeglasses, we compared the 3D profiles of a real and synthetic image pair (Fig. 9). In the profile of the real image, the intensities of reflected light around the eyes were almost 255, and the rest of the image had intensities near 150. On the other hand, in the profile of the synthetic image, most parts of the face had intensities near 255. From this observation, we expected that CycleGAN for our augmentation method was trained to remove the reflected light around eyes by increasing the overall intensities of the face rather than by adding information about the face to the areas of the reflected light.

Fig. 10 shows the failure cases of the proposed G2NG data augmentation. In the synthetic images without eyeglasses, black noise occurs around the eyes, and the eyes which are covered with the reflected lights are not realistically synthesized. However, the number of failure cases is much lower than that of the success cases. The numbers of the success and failed synthetic images are 32,992 and 4,191, respectively. Therefore, we can justify using CycleGAN for the proposed G2NG data augmentation.

2) QUANTITATIVE EVALUATION

Because it is not straightforward to quantitatively evaluate synthetically generated images, we assumed that if the synthetic images are realistic, the accuracy and validation rates of NIR FR would be increased in the G-NG positive FR scenarios after data augmentation. Therefore, as a quantitative evaluation of the proposed data augmentation, we compared the NIR FR validation rates with or without the use of the proposed data augmentation.

For this evaluation, instead of using LiNFNet, we utilized off-the-shelf DCNN architectures (Inception ResNet v1 [26] and VGGNet 16 [25]) to investigate the effects of the pro-

TABLE 9. The performance of NIR FR according to combinations of VGGNet 16_light, the depthwise separable convolution [18], and the linear bottleneck [21]. The details of the VGGNet 16_light architecture were already explained in Table 3. "DSC" and "LB" mean the depth separable convolution and linear bottleneck, respectively. The architectures are trained using Integrated NIR Face database.

	Accuracy (%)	Validation rate (%, FAR=0.1%)	# of parameters (million)	FLOPs (million)
Baseline (VGGNet 16_light)	97.47	88.96	14.85	148.39
Baseline + DSC	98.25	90.11	14.67	146.57
Baseline + LB	98.78	94.71	7.19	71.46
Baseline + DSC +LB (LiNFNet)	**99.44**	**97.67**	**7.03**	**70.12**

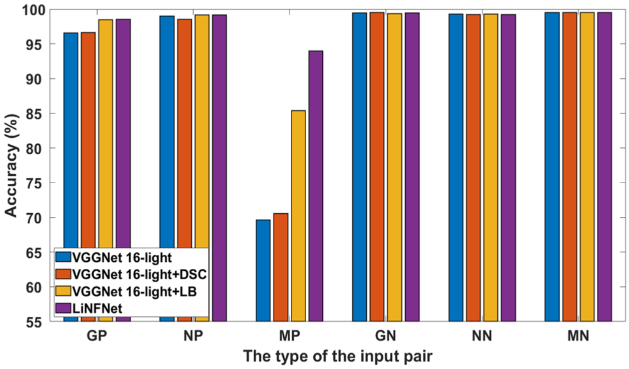


FIGURE 12. The NIR FR accuracy of the architectures in Table 9 according to the types of the input pair in Fig. 3. The abbreviations of the input pair types in X-axis have the same meanings as those in Fig. 4 and 5.

posed data augmentation. The results of the data augmentation in LiNFNet are discussed in Section VI-D.

We prepared several databases to train DCNN architectures (Table 6). The validation database was generated from CASIA NIR [46], and contains 2,000 pairs for each input pair type described in Fig. 3. The architectures were trained using fine-tuning [9], and the pretrained models were trained with data from the CASIA WebFace database [45].

The results of this experiment are shown in Fig. 11. For Inception ResNet v1 and VGGNet 16, the augmented training databases (CASIA VIS-NIR 2.0_AUG and PolyU-NIRFD_AUG) helped these architectures achieve higher validation rates of NIR FR than the original training databases (CASIA VIS-NIR 2.0 and PolyU-NIRFD). The augmented training databases considerably improved the accuracy of NIR FR for the mixed positive pairs (see Table 7). For Inception ResNet v1, the CASIA VIS-NIR 2.0_AUG and PolyU-NIRFD_AUG databases increased the NIR FR accuracy for the mixed positive pairs by 17.25% and 40.15%, respectively. In the case of VGGNet 16, the NIR FR accuracy for the mixed positive pairs increased by 16.25% and 30%, respectively. The use of the augmented training databases significantly improved the validation rate of the DCNN models for NIR FR in the G-NG positive FR scenarios.

B. DATABASES

In this section, we will explain the details of the training, validation, and test databases which were used in the experiments described in the next sections. As explained in Section III, the training stage consists of two steps: obtaining the pretrained model and fine-tuning.

TABLE 10. The RGB FR accuracy and validation rate for pretrained models of LiNFNet and the existing architectures [18], [19], [21], [22], [25], [26]. We use LFW database [52] as the validation database.

	Accuracy (%)	Validation rate (%, FAR=0.1%)
Inception ResNet v1 [26]	98.35	94.5
VGGNet 16 [25]	98.2	92.87
Xception [19]	98.27	93.46
1.0 MobileNet v1 [18]	96.92	82.93
1.0 MobileNet v2 [21]	97.43	91.43
ShuffleNet v2 (x2) [22]	97.9	91.54
LiNFNet	**98.53**	**94.6**

1) DATABASE FOR THE PRETRAINED MODEL

To obtain the pretrained models of LiNFNet and existing architectures [9], [18], [19], [21], [22], [25], [26], we utilized the CASIA WebFace database [45]. This database includes 453,414 RGB face images of 10,575 identities.

2) FINE-TUNING DATABASES FOR NIR FR

We prepared two fine-tuning databases for NIR FR: the Integrated NIR Face database and the Integrated NIR Face_AUG database. The Integrated NIR Face database was constructed by combining the CASIA VIS-NIR 2.0 [48] and PolyU-NIRFD [47] databases. This database includes 37,183 NIR face images for 948 identities. The Integrated NIR Face_AUG database is an augmented version of the Integrated NIR Face database; the database was constructed by CycleGAN-based G2NG data augmentation. When augmenting the database, we excluded the failure cases of the synthetic images shown in Fig. 10. This database contains 70,175 NIR face images for 948 identities. We did not follow the performance evaluation protocols of CASIA VIS-NIR 2.0, because these protocols are designed for heterogeneous FR (using both RGB and NIR face images).

Both databases were used to fine-tune LiNFNet and the existing architectures [9], [18], [19], [21], [22], [25], [26] in the experiments described in the following sections. These databases were used to show how the proposed data augmentation improves the accuracy and validation rate of NIR FR in the G-NG positive FR scenarios.

TABLE 11. The performances of LiNFNet and the existing architectures [18], [19], [21], [22], [25], [26] trained using Integrated NIR Face database.

	Accuracy (%)	Validation rate (%, FAR=0.1%)	# of parameters (Million)	FLOPs (Million)
Inception ResNet v1 [26]	97.39	87.37	22.93	228.95
VGGNet 16 [25]	97.9	91.57	14.91	149.05
Xception [19]	90.37	66.78	21.22	211.8
1.0 MobileNet v1 [18]	98.75	92.89	3.48	34.6
1.0 MobileNet v2 [21]	98.04	91.85	2.54	23.96
ShuffleNet v2 (x2) [22]	97.57	83.33	5.75	57.22
LiNFNet	**99.44**	**97.67**	**7.03**	**70.12**

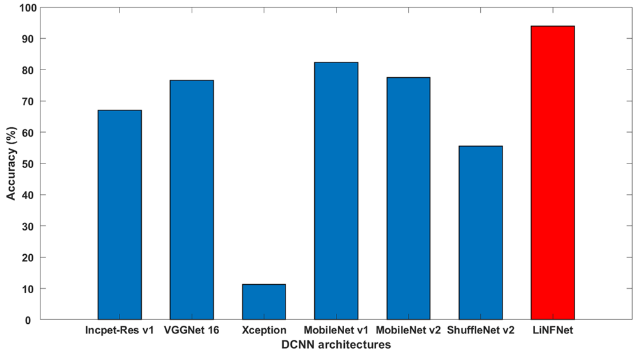


FIGURE 13. The NIR FR accuracy of LiNFNet and the existing architectures [18], [19], [21], [22], [25], [26] for the mixed positive pairs in Fig. 3 When the architectures are trained using the Integrated NIR Face database. Incept-Res v1 is an abbreviation of Inception ResNet v1.

TABLE 12. The NIR FR accuracy and validation rate of LiNFNet and the existing architectures [18], [19], [21], [22], [25], [26] trained using Integrated NIR Face_AUG database.

	Accuracy (%)	Validation rate (%, FAR=0.1%)
Inception ResNet v1 [26]	98.98	96.71
VGGNet 16 [25]	99.28	97.69
Xception [19]	97.01	88.43
1.0 MobileNet v1 [18]	98.96	97.44
1.0 MobileNet v2 [21]	98.91	96.81
ShuffleNet v2 (x2) [22]	98.71	94.37
LiNFNet	**99.54**	**99.09**

3) VALIDATION / TEST DATABASE

For the experiments described in the following sections, we used the CASIA NIR database [46] as the validation and test database, because this database has a number of G-NG mixed face classes including face images both with and without eyeglasses. By using the CASIA NIR database, we could construct a number of mixed positive pairs (Fig. 3 (c)) to evaluate the performance of the G-NG positive FR scenarios. The CASIA NIR database includes 3,938 NIR face images of 197 identities.

4) DATABASE CONFIGURATION

In the following sections, we report two experiments: the performance evaluation of LiNFNet, and the performance comparison of the proposed NIR FR system and existing NIR FR methods. We describe the database configuration for both experiments in Table 8. In these experiments, both the Integrated NIR Face and Integrated NIR Face_AUG databases were used as the training databases.

The CASIA NIR database [46], however, was utilized differently for two experiments. For the performance evaluation of LiNFNet, we acquired 12,000 pairs from the CASIA NIR database for validation; there are 2,000 pairs for each type of input pair (Fig. 3).

For the performance comparison of the proposed and existing NIR FR system, we assumed the identification scenarios. Therefore, we prepared the gallery and probe images using the CASIA NIR database. For the identification scenarios, we grouped the problems into two types: open-set and closed-set. The proposed system and Kim *et al.* [9] solve the open-set problem, and Zhang *et al.* [10] and Peng *et al.* [11] solve the closed-set problem.

C. TRAINING SETUP

In this section, we explain the detailed training settings for LiNFNet. The size of the NIR face images is 160 × 160 pixels. We conducted random crop and flip as the basic data augmentation apart from the proposed CycleGAN-based G2NG data augmentation. We set the iteration, batch size, and learning rate as 90,000, 32, and 0.001, respectively. Following the literature [2], we set the embedding size as 128. For all of the experiments in the following section, keep probability of dropout and weight decay were 0.8 and 0.00005, respectively, and we set center loss factor to 0.01 and center loss alpha to 0.9. When training LiNFNet, we used RMSProp, which is one of the gradient descent methods, and the fine-tuning method [9] was used as the training method. Ruder [53] has stated that RMSProp, Adadelt, and Adam are good gradient descent methods. Wilson *et al.* [54] also found that the image classification loss of RMSProp on the CIFAR dataset [55] was slightly lower than that of Adam. Since NIR FR is strongly associated with image classification, we chose RMSProp as the gradient descent method with which to train the DCNN architecture for NIR FR. We trained the LiNFNet architecture on a NVIDIA GTX 1080ti.

D. PERFORMANCE EVALUATION OF LINFNET

To evaluate the performance of the LiNFNet architecture in the G-NG positive FR scenarios, we conducted two exper-

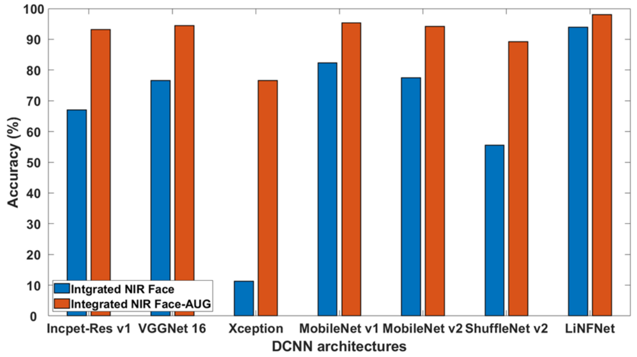


FIGURE 14. The NIR FR accuracy of LiNFNet and the existing archite-ctures [18], [19], [21], [22], [25], [26] for the mixed positive pairs according to the training databases.

TABLE 13. The identification rate of LiNFNet and the existing NIR FR methods [9]–[11] on G2NG test database in Table 8. "CDA" means the proposed CycleGAN-based G2NG data augmentation.

	Identification rate (%)
Zhang *et al.* [10]	71.56
Peng *et al.* [11]	85.94
Kim *et al.* (Inception ResNet v1) [9]	96.88
Kim *et al.* (VGGNet 16) [9]	95.63
Jo *et al.* [58]	99.06
LiNFNet (Ours)	**99.69**
LiNFNet + CDA (Ours)	**100**

iments. The first experiment was an ablation study of the LiNFNet architecture. We compared the performance of LiNFNet with existing DCNN architectures [18], [19], [21], [22], [25], [26] as the second experiment. As the performance metrics, we used accuracy, validation rate, the number of parameters, and FLOPs.

1) ABLATION STUDY

We conducted an ablation study to investigate the effect of the depthwise separable convolutions [18] and linear bottlenecks [21] in LiNFNet. For the baseline, we utilized VGGNet 16_light, a lightweight version of VGGNet 16. Using this baseline, we compared the performance of the following architectures: Baseline+DSC, Baseline+LB, and Baseline+DSC+LB (LiNFNet). The results of the experiment are summarized in Table 9.

The accuracy and validation rate of the Baseline+DSC were 0.8% and 1.2% higher than those of the baseline, respectively. Although the Baseline+DSC does not contribute much to the reduction of the number of parameters, this architecture reduces about 1.82 $\times$ 10^6 FLOPs over the baseline with respect to computational cost. Depth-wise separable convolution thus appears to be more suitable for the initial convolution layers of the VGGNet 16 architecture in NIR FR than the full convolution.

As shown in Table 9, the NIR FR accuracy and validation rate of the Baseline+LB increased by 1.3% and 5.8% over the baseline. This is because the linear bottlenecks extract better features for NIR FR by using a number of channels of the input activation than the full convolutions. In addition, the number of parameters and FLOPs of the Baseline+LB are about twice those of the baseline. Therefore, the linear bottleneck is the main factor in improving the performance of NIR FR in terms of accuracy, validation rate, memory, and computational complexity.

As shown in Table 9, the validation rate of LiNFNet increased over the baseline as much as the total increases of the Baseline+DSC and Baseline+LB. This means that the contributions of the two lightweight convolution modules (the depthwise separable convolution [18] and linear bottleneck [21]) to the improvement of the NIR FR validation rate do not overlap. Therefore, in order to construct the LiNFNet architecture, utilizing the lightweight convolution modules to VGGNet 16_light is extremely effective for improving the accuracy and validation rate of NIR FR. As shown in Fig. 12, LiNFNet showed considerable increase in NIR FR accuracy for the mixed positive pairs over other architectures. We demonstrated that LiNFNet is an efficient lightweight version of the VGGNet 16 architecture in the G-NG positive FR scenarios with respect to memory usage, computational complexity, and NIR FR accuracy.

2) COMPARISON WITH EXISTING DCNN ARCHITEC-TURES

In this section, we describe the performances of the following architectures as compared with LiNFNet: Inception ResNet v1 [26], VGGNet 16 [25], and the existing lightweight DCNN architectures [18], [19], [21], [22].

For Inception ResNet v1 [26], VGGNet 16 [25], MobileNet v1 [18], and MobileNet v2 [21], we used TensorFlow implementations. To evaluate the performance of Xception [19] and ShuffleNet v2 [22], we used the implementations reported in [50] and [51], respectively. For fair comparison, all of the architectures [18], [19], [21], [22], [25], [26] were modified to have a deep feature of 128 dimension. To do this, we replaced the fully connected layers of the existing architectures with that of the LiNFNet architecture shown in Table 4.

The first experiment was designed to evaluate the performances of the pretrained models of LiNFNet and other DCNN architectures [18], [19], [21], [22], [25], [26] in the RGB domain. The LFW database [52] was used as a validation database. The results of the experiment are summarized in Table 10. In general, the performance of a DCNN architecture decreased as the architecture became lighter. However, although LiNFNet is a lightweight version of VGGNet 16, LiNFNet had higher accuracy and validation rate than VGGNet 16, and also achieved the best performance amongst all architectures for the performance comparison.

For the second experiment, the performances of the architectures without the proposed G2NG data augmentation are summarized in Table 11. LiNFNet achieved the highest NIR FR accuracy and validation rate among all architectures described in Table 11. As shown in Fig. 13, LiNFNet had the best FR accuracy of the mixed positive pairs. Even though LiNFNet was trained without the proposed G2NG data augmentation, it could achieve a high accuracy of 94% in the G-NG positive FR scenario.

As shown in Table 10 and Fig. 13, The LiNFNet architecture is more effective at recognizing the mixed positive pairs in the NIR domain and the challenging face image pairs in the RGB domain than the existing DCNN architectures [18], [19], [21], [22], [25], [26]. In addition, LiNFNet has considerably fewer parameters and FLOPs than VGGNet 16. Although LiNFNet is slightly heavier than the existing lightweight architectures [18], [21], [22] described in Table 11, the accuracy and validation rate of LiNFNet are considerably higher than those of the competitors. Therefore, LiNFNet achieves a good balance between accuracy and computational complexity.

To explore the performance improvements achieved through the proposed data augmentation, all architectures [18], [19], [21], [22], [25], [26] were fine-tuned using the Integrated NIR Face_AUG database. The results of the performance evaluation are summarized in Table 12. After the proposed data augmentation, all of the architectures in Table 12 performed better than the no-augmentation versions shown in Table 11. From the results shown in Fig. 14, it is apparent that the proposed data augmentation is effective in improving accuracy for the mixed positive pairs. By integrating CycleGAN-based G2NG data augmentation and LiNFNet, the proposed NIR FR system achieved an accuracy and validation rate of more than 99%, and the proposed system also had a better ability to recognize the mixed positive pairs than the off-the-shelf DCNN architectures [18], [19], [21], [22], [25], [26].

E. PERFORMANCE COMPARISON OF THE PROPOSED NIR FR SYSTEM AND EXISTING METHODS

We compared the proposed system with the existing DCNN-based NIR FR methods [9]–[11], [58]. For this experiment, we reproduced the Zhang's method [10] and the Peng's method [11] known to have the NIR FR accuracies of around 98%. We verified that the two implemented methods achieved identification rates of 97.92% and 97.4%, respectively. These values are similar to those which are reported in [10] and [11]. Therefore, we verified that the implementations of [10] and [11] were correct. The work of Kim's method [9] and Jo's method [58] was also reproduced. Kim's method [9] achieved an identification rate of over 99%. The NIR FR method developed by Kim *et al.* [9] had a better ability to recognize the pairs that included only NIR face images without eyeglasses than the Zhang's method [10] and the Peng's method [11].

Despite the high reported accuracy of the existing NIR FR methods [9]–[11], [58], these results did not consider mixed positive pairs. Peng *et al.* [11] excluded NIR face images with eyeglasses in the training and test processes of FR, and Zhang *et al.* [10] utilized the PolyU-NIRFD database [47] as training and test databases to conduct performance evaluation; as shown in Table 1, there are few mixed positive pairs in the PolyU-NIRFD database. In the literature [9], an analysis of the G-NG positive FR scenarios was lacking. Jo and Kim [58] added simple reflected light patterns to the areas of the NIR face image around the eyes. However, the patterns did not prove to be the sufficiently realistic.

To compare the proposed NIR FR system with existing NIR FR methods [9]–[11], [58] in G-NG positive FR scenarios, we constructed a G2NG test database, as described in Table 8, and conducted performance evaluation of identification on the G2NG test database. The results of this experiment are presented in Table 13.

When using the proposed CycleGAN-based G2NG data augmentation to train the LiNFNet architecture, the identification rate of the architecture increased. The proposed data augmentation therefore contributes to an improvement in the identification rates on the G2NG test database. In addition, LiNFNet trained without CycleGAN-based data augmentation achieved 4% and 0.6% higher identification rates than Kim's method [9] and Jo's method [58], respectively. Therefore, the LiNFNet architecture itself is robust against reflected light in the G-NG positive FR scenarios. The proposed NIR FR system (LiNFNet + CDA) has the best NIR FR ability to recognize the mixed positive pairs among the NIR FR methods, as shown in Table 13.

VII. CONCLUSION

In this paper, we propose a DCNN-based fast NIR FR system robust to reflected light. The proposed system has two contributions: one is the CycleGAN-based G2NG data augmentation, and the other is LiNFNet. Through these two contributions, the performance of the proposed NIR FR system is improved with respect to accuracy and computational complexity. Especially, the proposed NIR FR system considerably improves the accuracy of DCNN-based NIR FR in G-NG positive FR scenarios. We showed that the proposed system has advantages in terms of striking a balance between accuracy and the computational complexity of NIR FR over existing lightweight architectures [18], [19], [21], [22] as well as off-the-shelf DCNN architectures [25], [26]. The proposed system also has the best identification rate, compared to the existing NIR FR methods [9]-[11], on the G2NG test database, which includes mixed positive pairs, as shown in Fig. 3. The system achieved an identification rate of 100% on the G2NG test database.

Before discussing future works, it is worth mentioning the pros and cons of our NIR FR system compared to existing methods [56]–[58]. Based on the experiment of [9], the proposed NIR FR method is expected to have an advantage over existing RGB FR methods [56], [57] regarding FR validation rate under poor lighting condition. However, the architecture of Wu *et al.* [56] can be more versatile than LiNFNet for different modalities of FR, because it was designed to solve not only RGB FR scenarios, but also infrared-visible heterogeneous FR scenarios. As compared to the method of Jo *et al.* [58], the proposed system has a better FR validation rate than the competitor; however, DCNN architecture used in [58] is less complex than LiNFNet.

Based on the pros and cons of the proposed system, we can set two possible future directions of research: 1. Improving LiNFNet to handle various modalities of FR, 2. Developing a DCNN architecture which can produce more efficient facial representations than LiNFNet.

Also, the accuracy and validation rate of NIR FR depend upon the contents and characteristics of the training and validation databases. To address this problem, we will research methods that reduce the sensor dependency of NIR FR.

REFERENCES

[1] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Swansea, U.K., 2015, pp. 1–12.

[2] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 815–823.

[3] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 2892–2900.

[4] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face recognition with very deep neural networks," 2015, *arXiv:1502.00873*. [Online]. Available: http://arxiv.org/abs/1502.00873

[5] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6738–6746.

[6] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5265–5274.

[7] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1701–1708.

[8] M. A. Abuzneid and A. Mahmood, "Enhanced human face recognition using LBPH descriptor, multi-KNN, and back-propagation neural network," *IEEE Access*, vol. 6, pp. 20641–20651, 2018, doi: 10.1109/ACCESS.2018.2825310.

[9] J. Kim, H. Jo, M. Ra, and W.-Y. Kim, "Fine-tuning approach to NIR face recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 2337–2341.

[10] X. Zhang, M. Peng, and T. Chen, "Face recognition from near-infrared images with convolutional neural network," in *Proc. 8th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Yangzhou, China, Oct. 2016, pp. 13–15.

[11] M. Peng, C. Wang, T. Chen, and G. Liu, "NIRFaceNet: A convolutional neural network for near-infrared face identification," *Information*, vol. 7, no. 4, pp. 61–74, Oct. 2016, doi: 10.3390/info7040061.

[12] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[13] A. D. Pozzolo, O. Caelen, and G. Bontempi, "When is undersampling effective in unbalanced classification tasks?" in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Porto, Portugal, 2015, pp. 200–215.

[14] S.-J. Yen and Y.-S. Lee, "Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset," in *Intelligent Control and Automation*. Berlin, Germany: Springer, 2006, pp. 731–740.

[15] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newslett.*, vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002, doi: 10.1613/jair.953.

[17] Q. Wang, X. Zhou, C. Wang, Z. Liu, J. Huang, Y. Zhou, C. Li, H. Zhuang, and J.-Z. Cheng, "WGAN-based synthetic minority over-sampling technique: Improving semantic fine-grained classification for lung nodules in CT images," *IEEE Access*, vol. 7, pp. 18450–18463, 2019, doi: 10.1109/ACCESS.2019.2896409.

[18] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861

[19] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1251–1258.

[20] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," 2016, *arXiv:1602.07360*. [Online]. Available: http://arxiv.org/abs/1602.07360

[21] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 4510–4520.

[22] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 116–131.

[23] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6848–6856.

[24] D. Lai, X. Zhang, Y. Bu, Y. Su, and C.-S. Ma, "An automatic system for real-time identifying atrial fibrillation by using a lightweight convolutional neural network," *IEEE Access*, vol. 7, pp. 130074–130084, 2019, doi: 10.1109/ACCESS.2019.2939822.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015, pp. 1–14.

[26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 13th AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, 2016, pp. 4278–4284.

[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[28] A. Kirzhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2818–2826.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[32] S. Ma, J. Fu, C. W. Chen, and T. Mei, "DA-GAN: Instance-level image translation by deep attention generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5657–5666.

[33] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," 2017, *arXiv:1711.04340*. [Online]. Available: http://arxiv.org/abs/1711

[34] S.-W. Huang, C.-T. Lin, S.-P. Chen, Y.-Y. Wu, P.-H. Hsu, and S.-H. Lai, "AugGAN: Cross domain adaptation with GAN-based data augmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 718–731.

[35] Y. Shen, P. Luo, P. Luo, J. Yan, X. Wang, and X. Tang, "FaceID-GAN: Learning a symmetry three-player GAN for identity-preserving face synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 821–830.

[36] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Cambridge, MA, USA, 2014, pp. 2672–2680.

[37] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, ''Least squares generative adversarial networks,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2794–2802.

[38] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, ''Stacked generative adversarial networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5077–5086.

[39] A. Ghosh, V. Kulharia, V. Namboodiri, P. H. S. Torr, and P. K. Dokania, ''Multi-agent diverse generative adversarial networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8513–8521.

[40] S. Gurumurthy, R. K. Sarvadevabhatla, and R. V. Babu, ''DeLiGAN: Generative adversarial networks for diverse and limited data,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 166–174.

[41] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, ''Unsupervised pixel-level domain adaptation with generative adversarial networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3722–3731.

[42] B. M. Lake, R. Salakhutdinov, and J. B. Tenebaum, ''Human-level concept learning through probabilistic program induction,'' *Science*, vol. 350, no. 6266, pp. 1332–1338, Dec. 2015, doi: 10.1126/scien-ce.aab3050.

[43] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, ''EMNIST: An extension of MNIST to handwritten letters,'' 2017, *arXiv:1702.05373*. [Online]. Available: http://arxiv.org/abs/1702.05373

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, ''Unpaired image-to-image translation using cycle-consistent adversarial networks,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[45] D. Yi, Z. Lei, S. Liao, and S. Z. Li, ''Learning face representation from scratch,'' 2014, *arXiv:1411.7923*. [Online]. Available: http://arxiv.org/abs/1411.7923

[46] S. Z. Li, R. Chu, S. Liao, and L. Zhang, ''Illumination invariant face recognition using near-infrared images,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 627–639, Apr. 2007, doi: 10.1109/TPAMI.2007.1014.

[47] B. Zhang, L. Zhang, D. Zhang, and L. Shen, ''Directional binary code with application to PolyU near-infrared face database,'' *Pattern Recognit. Lett.*, vol. 31, no. 14, pp. 2337–2344, Oct. 2010, doi: 10.1016/j.patrec.2010.07.006.

[48] S. Z. Li, D. Yi, Z. Lei, and S. Liao, ''The CASIA NIR-VIS 2.0 face database,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Portland, OR, USA, Jun. 2013, pp. 348–353.

[49] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, ''Image-to-Image translation with conditional adversarial networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1125–1134.

[50] K. S. Lee. *TensorFlow Implementation of the Xception Model by François Chollet*. Accessed: Jun. 2019. [Online]. Available: https://github.com/kwotsin/TensorFlow-Xception

[51] C.-T. Ho. *Tensorflow ShuffleNet v2 Implementation*. Accessed: Oct. 2019. [Online]. Available: https://github.com/timctho/shufflenet-v2-tensorflow

[52] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, ''Labeled faces in the wild: A database forstudying face recognition in unconstrained environments,'' Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.

[53] S. Ruder, ''An overview of gradient descent optimization algorithms,'' 2016, *arXiv:1609.04747*. [Online]. Available: http://arxiv.org/abs/1609.04747

[54] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, ''The marginal value of adaptive gradient methods in machine learning,'' in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, 2017, pp. 4148–4158.

[55] A. Krizhevsky, ''Learning multiple layers of features from tiny images,'' M.S. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2009.

[56] X. Wu, R. He, Z. Sun, and T. Tan, ''A light CNN for deep face representation with noisy labels,'' *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018, doi: 10.1109/TIFS.2018.2833032.

[57] H. H. Zheng and Y. X. Zu, ''A normalized light CNN for face recognition,'' *J. Phys., Conf. Ser.*, vol. 1087, Sep. 2018, Art. no. 062015, doi: 10.1088/1742-6596/1087/6/062015.

[58] K. Jo, ''NIR reflection augmentation for DeepLearning-based NIR face recognition,'' *Symmetry*, vol. 11, no. 10, p. 1234, Oct. 2019, doi: 10.3390/sym11101234.

JEYEON KIM received the B.S. degree in electronic engineering from Hanyang University, Seoul, South Korea, in 2015, where he is currently pursuing the Ph.D. degree in electronics and computer engineering. His research interests include face recognition, lightweight deep learning architecture, machine learning, image-to-image translation, and camera geometry.

MOONSOO RA received the B.S. and Ph.D. degrees in electronics and computer engineering from Hanyang University, Seoul, South Korea, in 2011 and 2019, respectively. He is currently working as a Founding Member and the CTO of LightVision Inc. His research interests include pattern recognition, machine learning, autonomous vehicle, and video surveillance.

WHOI-YUL KIM received the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1989. From 1989 to 1994, he was an Assistant Professor with The University of Texas at Dallas. He joined Hanyang University, in 1994, where he is currently a Professor with the Department of Electronic Engineering. His research interests include health monitoring using computer vision, intelligent surveillance, machine vision, advanced driver assistance systems, and 3-D vision systems in sports.