IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 8, 2021, accepted March 21, 2021, date of publication March 24, 2021, date of current version April 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068517

Dog Nose-Print Identification Using Deep Neural Networks

HAN BYEOL BAE[1], DAEHYUN PAK[2], AND SANGYOUN LEE[1], (Member, IEEE)
[1]Department of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea
[2]Petnow Inc., Seoul 06765, South Korea

Corresponding author: Sangyoun Lee (syleee@yonsei.ac.kr)

This work was supported by the Research and Development Program for Advanced Integrated-Intelligence for Identification (AIID) through the National Research Foundation of Korea (NRF) funded by Ministry of Science and ICT under Grant NRF-2018M3E3A1057289.

ABSTRACT Recently, there has been rapid growth in the number of people who own companion pets (cats and dogs) due to low birth rates, an increasingly aging population, and an increasing number of single-person households. This trend has resulted in a growing interest in problems requiring solutions, such as missing pets and false insurance claims. Traditional non-biometric-based methods cannot address these problems. This paper proposes a novel deep-learning model that can extract discriminative features through dog nose-print patterns for individual identification. We present a robust baseline for how individual dogs can be identified. The proposed dog nose network (DNNet) is a convolutional neural network (CNN)-based Siamese network structure comprising feature extraction and self-attention modules. Moreover, there is no need for a separate scanning device because it uses popular mobile devices to acquire the dataset. Besides high recognition performance, the proposed method also ensures simplicity and efficiency. The proposed method achieves better recognition performance than state-of-the-art methods for the collected dog nose-print dataset. It achieves recognition performance superior to state-of-the-art methods for the collected dog nose-print dataset. Using multiple datasets through cross-validation, we acquired an average identification accuracy of 98.972% with the Rank-1 approach. Additional performance benefits were demonstrated through the receiver operating characteristic (ROC) curve, t-distributed stochastic neighbor embedding (t-SNE), and confusion matrix.

INDEX TERMS Dog identification, Siamese network, convolutional neural network, residual learning, attention mechanism.

I. INTRODUCTION

Animal biometrics has been a promising area of study in the fields of computer vision and machine learning in recent years. It involves extracting discriminative features by considering morphological or biometric traits, such as visual appearance, facial features, coat patterns, and nose-print patterns [1]–[3]. Accordingly, animal-biometric-based identification systems have been applied in various areas for animal identification, management, and behavioral analyses. Animals, especially cats and dogs, are common companion pets in our society and have shared a familiar environment with humans for a long time. The harmonious coexistence between people and animals and the associated responsibility of owning and raising a pet must be considered from the perspective that companion pets are not a hobby but an essential culture in modern society. For example, effective registration and management of companion pets require handling insurance frauds and prompt handling of missing companion pets. Therefore, the animal-biometric-based identification system is a vital tool for managing and monitoring companion pets. The number of incidents associated with missing animals can be significantly minimized through identification and tracking by clearly connecting the owners and pets. Moreover, by enabling successful data registration, valuable data can be collected to overcome the limitations imposed by insufficient datasets.

Traditionally, non-biometric methods use intuitive and physical forms and are classified into three categories as shown in Fig. 1: permanent methods (ear tipping, ear notches, tattoos, microchip implant, and freeze branding), semi-permanent methods (ear tags, collar tags), and

The associate editor coordinating the review of this manuscript and approving it for publication was Chang-Hwan Son.

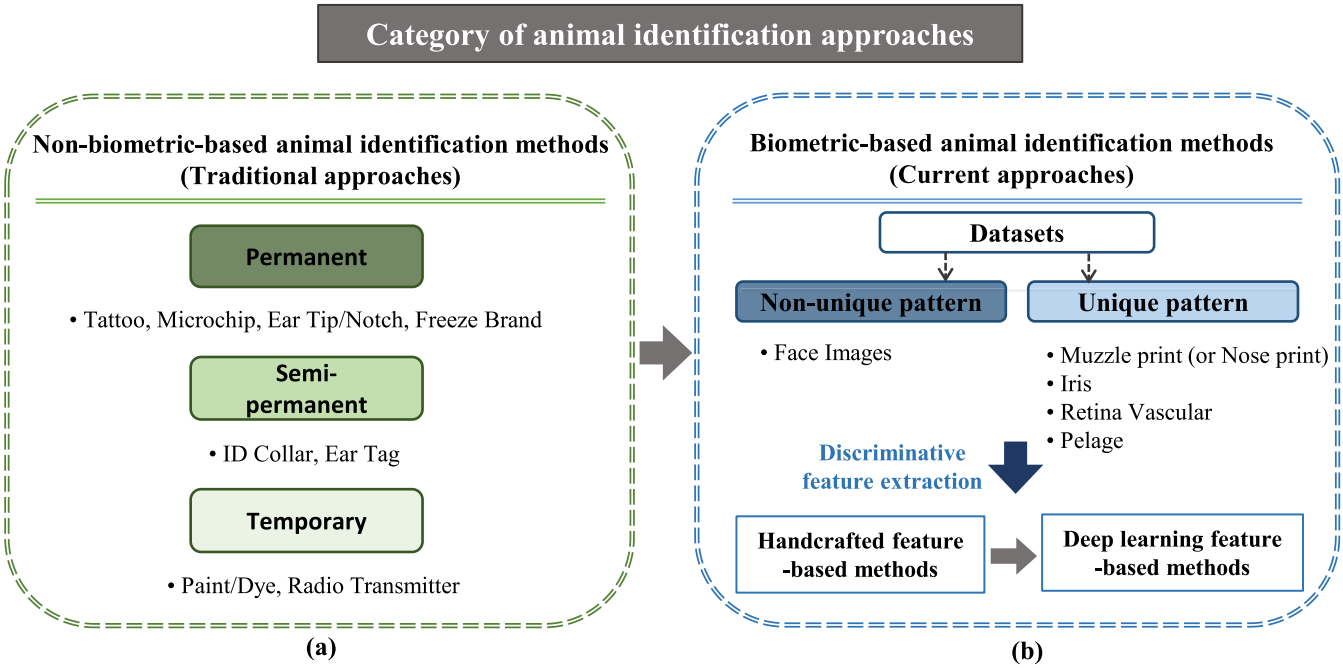

FIGURE 1. Flow chart for categorization of animal identification approaches: (a) non-biometric-based and (b) biometric-based animal identification methods.

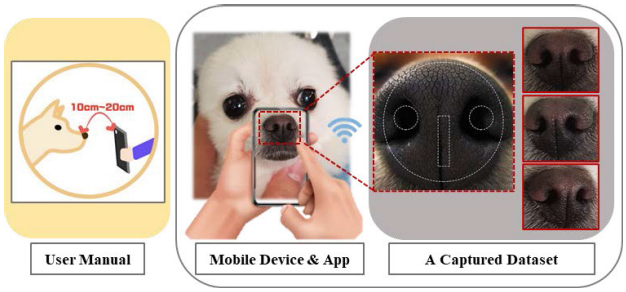

FIGURE 2. Dataset collection process.

temporary methods (paint or dye, radio-frequency-based identification [RFID], global positioning system [GPS] trackers) [3], [4]. Such non-biometric identification techniques use chip implants, deformation of skin tissues, and the wearing of specific devices. However, these methods can cause considerable pain to the animals, and concerns have been noted for tag loss or fraud and animal-welfare problems. Therefore, biometric-based identification methods are becoming popular as alternatives to the existing non-biometric identification methods. As shown in Fig. 1, biometric information, such as muzzle print (or nose-print), iris, retinal vessel, pelage pattern, and facial images, are used as the basis for identification in biometric-based identification systems. Some studies [5]–[16] have considered shortcomings in animal identification based on facial features. Face images are the most commonly used biometric research tools, both in animals and humans, and datasets can be easily collected through various media. However, animal faces are affected by various lighting changes, poses, and large-scale textural changes. Consequently, many studies have used unique patterns of the animal body parts for identification. These unique patterns remain constant irrespective of the age of the animal and can contribute discriminative features. Among these unique patterns, the most studied is the muzzle print (or nose-print) pattern, which can be used similar to human fingerprints [3], [4], [17]–[27]. Coldea [17] described the need for unique nose-print patterns to identify dogs. Existing animal identification systems use handcrafted features to identify unique discriminative features in the animals. However, the handcrafted features acquired in an open environment without constraints pose challenges to extract discriminative features. The deep-learning approach has recently garnered much attention for identifying species or individual animals using deep features.

This study proposes a novel dog-nose network (DNNet) framework based on deep learning to enhance the identification of individual dogs. As illustrated in Fig. 4, the proposed DNNet follows the Siamese network [28] structure based on a convolutional neural network (CNN) to identify discriminative features with the dog nose-print patterns. As shown in Fig. 3, each DNNet of the Siamese network involves two-step feature extraction and attention modules. First, the feature extraction module applies a deep residual network [29] as a backbone model, after which the additional layers are added to lower the feature map channels. Second, the attention module is aimed at obtaining superior distinctive traits by applying a non-local (NL) self-attention

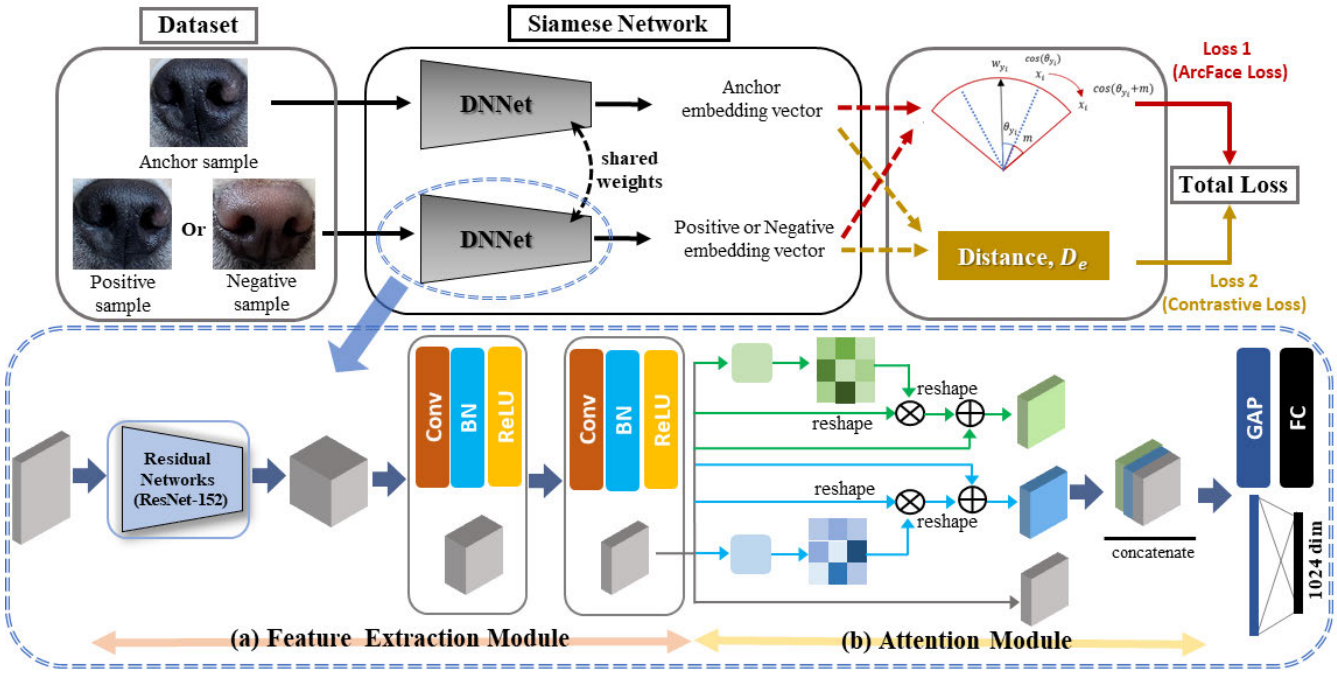

FIGURE 3. Flow chart of DNNet framework.

mechanism [30], which simultaneously considers the channel and spatial attention of the feature extraction module's feature map.

The original feature maps obtained through the first step are then concatenated with the channel axis for each map obtained through the channel and spatial attention module in the second step. The final embedding vector is obtained through a fully connected (FC) layer. We used contrastive loss [31] to optimize the DNNet because it can widen the inter-class distances and narrow the intra-class distances. The contrastive loss is calculated by checking and applying the binary label to a pair of positive-negative inputs. We also added additional margin-based loss (ArcFace) [32] to extract the discriminative embedding vectors of the DNNet. The ArcFace loss is considered with the contrastive loss to optimize DNNet. The experimental outcomes indicate that the proposed framework illustrates superior recognition performance to state-of-the-art methods for the collected dog nose-print dataset.

The contributions of our proposed framework are as follows:

- The proposed DNNet method improves individual identification systems' performance through nose-print patterns based on deep learning techniques. Our method is the first attempt to identify an individual dog's nose-print patterns based on deep learning models. We provide a robust baseline model through the DNNet method for individual identification systems.
- We ensure stable and discriminative feature extraction by integrating the DNNet modules into end-to-end training and combined objective functions to optimize the network.
- We experimentally demonstrate the superior performance for our collected dog nose-print dataset compared to state-of-the-art methods. We acquired an average identification accuracy of 98.972% with the Rank-1 approach. Additional performance benefits were demonstrated through the receiver operating characteristic ROC) curve, t-distributed stochastic neighbor embedding (t-SNE), and confusion matrix.

The remainder of this paper is organized as follows. Section II provides a review of studies related to animal classification systems. Section III presents a detailed explanation of the proposed framework, including the network architecture, obtaining a discriminative embedding vector, and enhancing performance. Section IV describes the experimental setup and dataset and presents the analysis results of the experiments. Finally, the conclusions are described in Section V.

II. RELATED WORK

Traditionally, approaches to identifying animals have adopted non-biometric-based permanent, semi-permanent, and temporary methods. However, these non-biometric-based methods incur additional cost resulting from separate labor. Moreover, for animal tags, duplication and forgery are possible, and animals may also be subjected to mental and physical pain resulting from stimulation and deformation inflicted on their bodies. Thus, biometric-based methods have become popular as alternatives to individual animal identification systems for effective and stable performance.

This section briefly reviews the following biometric-based identification methods: handcrafted feature-based and deep-learning-based.

A. HANDCRAFTED FEATURE-BASED METHODS

Kumar *et al.* [6] proposed a method of identifying cattle using face images. This method used the AdaBoost detection algorithm to extract feature vectors via conventional machine learning methods, such as principal component analysis (PCA), linear discriminative analysis (LDA), and independent component analysis (ICA) from the cropped face images. Matkowski *et al.* [10] suggested the use of panda face images to connect the texture-based local binary pattern (LBP) features and Gabor features. Crowe *et al.* [15] proposed an identification system for face images using the normalized multiscale LBP (MLBP) features. However, animal face images are prone to changes in texture, lighting, and pose. Many studies have focused on distinct patterns, such as muzzle print or nose-print, and these unique patterns are popular in biometric-based identification systems because they are not altered with age, similar to human fingerprints, and represent the unique features of an animal. Some studies [3], [4], [20]–[23], [26], [27] suggest applying various handcrafted feature-based methods to cattle using the muzzle print as the feature. Taha *et al.* [19] presented an identification system for Arabian horses using muzzle print images, with three steps: scale-invariant feature transform (SIFT) extraction, SIFT matching, and random sample consensus (RANSAC) optimization. Chen *et al.* [24] proposed an identification method for cats using nose-print patterns. They prevented low-quality image problems caused by the use of separate scanning equipment to capture nose patterns by applying sparse representation features and a support vector machine (SVM) classifier. Chakraborty *et al.* [25] used cropped muzzle images of pigs for breed identification; their system involved feature spaces of each of the four pig breeds via gradient significance map (GSM) and maximal likelihood (ML) estimation. Chehrsimin *et al.* [33] considered individual identification via unique pelage patterns of the Saimaa ringed seal. Segmentation and postprocessing were performed to identify target parts. However, these handcrafted feature-based approaches do not guarantee high-performance outcomes, rely primarily on datasets, and require extensive preprocessing depending on the environmental impacts.

B. DEEP LEARNING FEATURE-BASED METHODS

Recently, deep learning has become a key area of development in computer vision and is a vital part of cutting-edge technology. Deep-learning approaches are popular for the recognition, classification, detection, and tracking of objects. Therefore, animal species or individual identification recognition through deep learning is gradually gaining attention. The CNN is a popular deep-learning architecture that has demonstrated outstanding performance in various computer vision tasks [29], [34]–[37]. Hansen *et al.* [13] proposed an identification system of individual livestock with pig face images that uses a CNN model for training with an artificially augmented dataset from an unconstrained commercial farm environment. Deb *et al.* [14] presented a face recognition system called PrimNet, where mobile applications were used to directly obtain images of three primates in the wild: lemurs, golden monkeys, and chimpanzees. Hou *et al.* [16] used CNN with deep learning to propose a new individual identification system for the giant panda; they ensured the effectiveness and reliability of the identification model by considering multiple treatments under various conditions, such as large face angle, low brightness, and high saturation. Wang *et al.* [7] used a CNN with residual learning to study the unique facial features of the panda for gender classification. Kumar *et al.* [3] proposed an approach using deep-learning architectures, such as a CNN and a deep belief network (DBN), for individual cattle identification. The performance of this approach was superior to that of the handcrafted feature-based approach that was previously applied using muzzle print images. Favorskaya and Pakhirka [38] presented animal species identification in the wildlife based on muzzle and shape features using a joint CNN. Hu *et al.* [39] proposed a cow-identification system based on the fusion of deep parts features; they use side-view images, including the head, trunk, and leg parts of the cow, to identify individual cows.

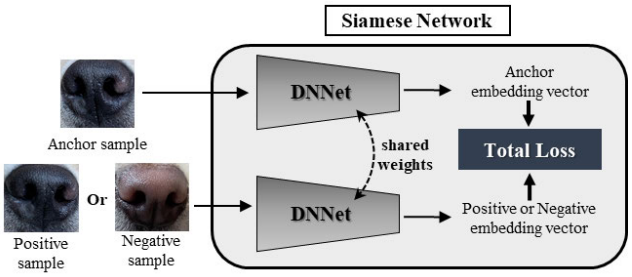

FIGURE 4. Proposed siamese network structure; weights of each network in siamese network are shared.

III. PROPOSED METHOD

This section presents the proposed DNNet framework that enhances the dog-identification system performance for the collected dog nose-print dataset. We first explain the general overview and then present the detailed DNNet modules.

A. BASELINE OVERVIEW

The aim of our proposed framework is to determine a biometric-based individual identification system that can extract discriminative features using unique patterns of dog nose-prints, as illustrated in Fig. 3. Because there are no available public dog nose-print datasets, we collected a dog nose-print dataset using mobile devices, as shown in Fig. 2. The proposed framework uses the Siamese network structure, where the primary aim is to solve the verification problem in [28]. Each DNNet that forms the Siamese network of the proposed framework shares the weights with the other networks, as illustrated in Fig. 4. The DNNet includes two steps:

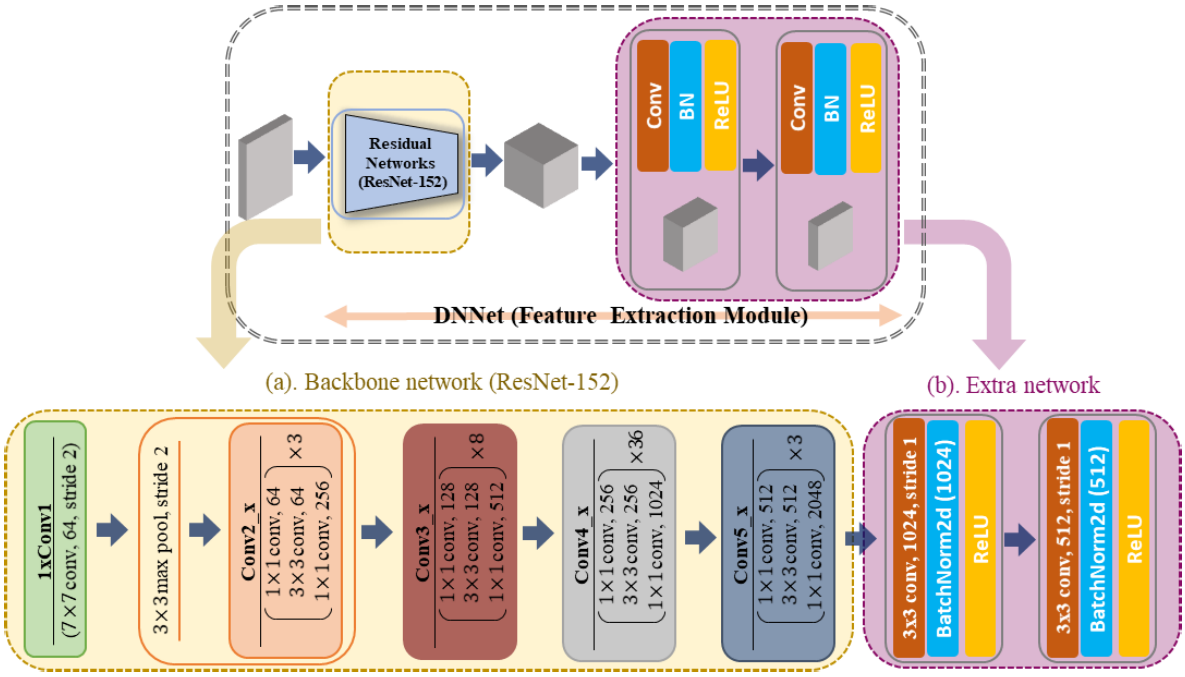

FIGURE 5. DNNet feature extraction module, consisting of backbone network (ResNet-152) and extra network.

feature extraction and attention modules. As shown in Fig. 5, in the first step, the feature extraction module is applied to the deep residual network [29] as a CNN-based backbone network. ResNet-152 is used here as the deep residual network, except for the last average pooling layer and FC layer. After performing the backbone network, we create two more building blocks to minimize the feature map channels. These building blocks have a convolution layer, batch normalization (BN) [40], and a ReLU [41] activation function, as shown in Fig. 5. The attention module is the second step in the proposed DNNet framework, as illustrated in Fig. 3. It enhances the output feature maps from the feature extraction module by applying an NL-based self-attention system [30] that aids in capturing the correlations between both channel and spatial attention around the original feature maps, as shown in Fig. 7. The original feature map obtained through the first step is concatenated to the channel axis for each feature map acquired through the channel and spatial attention module in the second step. The concatenated feature maps are passed through the global average pooling (GAP) and FC layers to obtain the final 1,024-dimensional embedding vector. The embedding vectors extracted from each branch of the Siamese network structure based on the proposed DNNet are used to calculate two objective functions to optimize the network. The input anchor image and corresponding positive or negative paired images are always considered together. In the contrastive loss, the binary label functions as the determinant of whether the relationship between the input anchor image and the corresponding paired image is positive or negative. The purpose of the contrastive loss is to widen the distance between the various classes and narrow the distances within each class. Moreover, the ArcFace loss is calculated by remeasurement, based on the input and paired data together with the label information. The ArcFace loss maximizes the decision boundaries through margin-based representation in an angular space to acquire discriminative features. We always consider both these losses simultaneously, and each loss uses a different optimization. Thus, discriminative and stable embedding vectors are obtained through these integrated modules of the DNNet into end-to-end training and combined objective functions to optimize the network.

B. FEATURE EXTRACTION MODULE

The feature extraction module is the first step in the DNNet framework and uses the deep residual network as the backbone model to create additional networks behind the backbone network. As shown in Fig. 5, the backbone network follows the ResNet-152, except for the last GAP and FC layers. The residual network used here, ResNet, was that proposed by He *et al.* [29]. The residual network presents a solution to the degradation problem, in which accuracy saturates as network depth increases and then rapidly degrades. ResNet-152 is layered with building blocks as a bottleneck design, and each building block is highly relevant as a residual unit, as illustrated in Fig. 6. Each residual unit comprises convolution layers, BN [40], and ReLU [41] activation functions, defined by the following equation:

$$y = F(x, \{W_i\}) + x, \tag{1}$$

where x and y are the input and output vectors of the residual unit, respectively. F represents the residual mapping function, and W_i is the weight corresponding to the residual function.

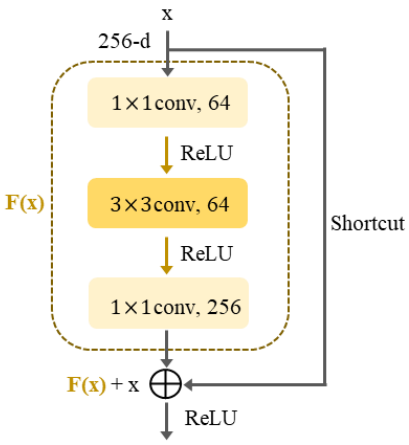

FIGURE 6. Sample of residual unit (ResNet-152).

$y = F(x, \{W_i\}) + x$ indicates that the input vector x is added to residual mapping function F according to the shortcut connection.

Based on the identity mapping function $H(x)$ as an example, it is fitted by several stacked layers for input x. Therefore, rather than expecting the stacked layers to be approximately $H(x)$, the learned layers will be close to the residual mapping function $F(x) := H(x) - x$. The shortcut connections in Eq. (1) do not increase the network's additional parameters or computational complexity. If the channel dimensions x and F do not match, a linear projection W_s is applied to the shortcut connections to match the dimensions. W_s is used only for dimension matching and is based on the following equation:

$$y = F(x, \{W_i\}) + W_s x. \tag{2}$$

As illustrated in Fig. 5, ResNet-152 is used, except for the GAP and FC layers, and is divided into five parts. Conv1 includes a convolution layer with a 7×7 convolution kernel. The Conv i_x $(i = 2, 3, 4, 5)$ building blocks as a bottleneck design consists of 3, 8, 36, and 3 residual units, respectively. As shown in Fig. 6, the structure of the residual units has three layers. The first and third layers are 1×1 convolution filters, and the second layer is a 3×3 convolution filter. The Input and output of the residual unit channel dimensions are matched by changing the number of 1×1 convolution filters.

After obtaining the output feature maps of the backbone network, we added additional networks to reduce the feature map channels. The reason for this channel reduction is to ensure superior feature aggregation instead of high complexity due to excess channels in the second step of the proposed framework, i.e., the attention module. As shown in Fig. 5, the additional network has two blocks, and each block has a convolution layer, BN, and a ReLU activation function.

C. ATTENTION MODULE

The attention module is the second step of the proposed DNNet framework, and the attention mechanism aims to identify the most informative components that control unnecessary elements in the feature map of the input image and to focus on the discriminative features. SENet [42] was proposed as an efficient method for learning channel attention for inter-channel correlation of the convolutional features. CBAM [43] presents both channel and spatial attention methods through average and max pooling along with several convolution layers. Wang *et al.* [44] proposed an NL-based self-attention model for video classification. The A^2-Nets [45] method proposed a double attention block to determine the novel relation features from the spatial-temporal spaces of the images. Lin *et al.* [46] proposed a novel framework containing sequential dual attention block (SDAB) for removing rain streaks in a single image. We applied the dual attention network (DAN) [30] as the second step in the DNNet framework; the DAN presents NL-based spatial and channel attention to informational features around feature maps. As shown in Fig. 7, the DAN channel and spatial attention are applied to the output feature maps from the feature extraction module of the DNNet. A detailed description of this module is as follows.

First, the structure of the channel attention module is shown in Fig. 7(a). We directly compute the channel attention map $\mathbf{X} \in \mathbb{R}^{C \times C}$ from input feature maps $\mathbf{A} \in \mathbb{R}^{C \times H \times W}$, where C is the number of the channel of input feature maps, and $H \times W$ is the size of the input feature map. We reshape $\mathbf{A}$ to $\mathbb{R}^{C \times N}$, and perform matrix multiplication between $\mathbf{A}$ and $\mathbf{A^T}$. We then obtain the channel attention map $\mathbf{X} \in \mathbb{R}^{C \times C}$ through a softmax layer:

$$x_{ji} = \frac{exp(A_i \times A_j)}{\sum_{i=1}^{C} exp(A_i \times A_j)}, \tag{3}$$

where x_{ji} is the i^{th} channel's influence on the j^{th} channel. Then, the outcome of the matrix multiplication between $\mathbf{X^T}$ and $\mathbf{A}$ is reshaped into $\mathbb{R}^{C \times H \times W}$. Finally, we multiply the reshaped result by a scale parameter β and perform an element-wise summation operation with the input feature map $\mathbf{A}$ to acquire the final channel attention map $\mathbf{E} \in \mathbb{R}^{C \times H \times W}$:

$$E_j = \beta \sum_{i=1}^{C} (x_{ji} A_i) + A_j, \tag{4}$$

where β is initialized as 0 and learns more weight gradually [47]. As shown in Eq. (4), the final channel attention map $\mathbf{E}$ includes the weighted sum for all channel features and can describe the long-term dependencies between the feature maps to boost the discriminant features.

Second, the structure of the spatial attention module is shown in Fig. 7(b). Given an input feature map $\mathbf{A} \in \mathbb{R}^{C \times H \times W}$, it is fed to two convolution layers to obtain new feature maps $\mathbf{B}$ and $\mathbf{C}$, where $\{\mathbf{B}, \mathbf{C}\} \in \mathbb{R}^{C \times H \times W}$. Then, $\mathbf{B}$ and $\mathbf{C}$ are reshaped into $\mathbb{R}^{C \times N}$, where $N = H \times W$ is the number of pixels. Thereafter, we perform matrix multiplication between the transpose of $\mathbf{B}$ and $\mathbf{C}$, and a softmax layer is applied to calculate the spatial attention map $\mathbf{S} \in \mathbb{R}^{N \times N}$:

$$s_{ji} = \frac{exp(B_i \times C_j)}{\sum_{i=1}^{N} exp(B_i \times C_j)}, \tag{5}$$

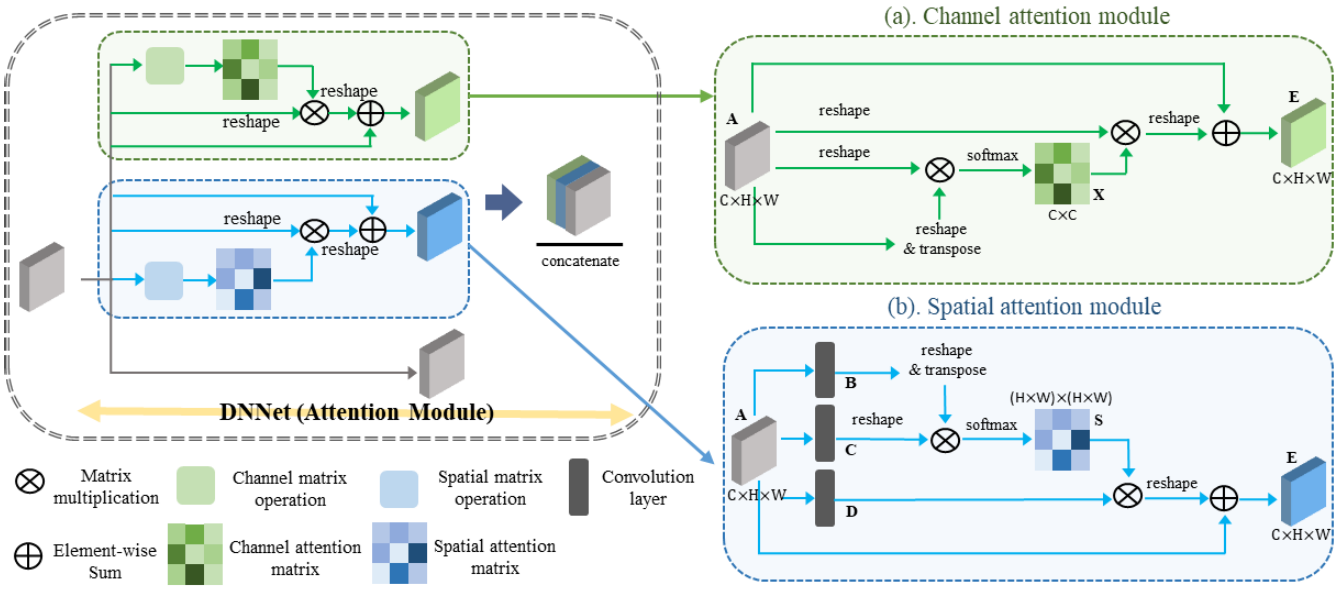

FIGURE 7. Attention module of DNNet: (a) Channel and (b) Spatial attention modules.

where s_{ji} is a measure of the impact of the i^{th} pixel on the j^{th} pixel. Closer feature representations of the two pixels result in stronger correlations between them. Next, we feed the input feature map $\mathbf{A}$ to a convolution layer to obtain a new feature map $\mathbf{D} \in \mathbb{R}^{C\times H\times W}$, which is reshaped to $\mathbb{R}^{C\times N}$. Then, we perform matrix multiplication between $\mathbf{D}$ and $\mathbf{S}^{\mathbf{T}}$, and the result is reshaped into $\mathbb{R}^{C\times H\times W}$. Finally, we multiply the reshaped result by a scale parameter α and perform an element-wise summation with the input feature map $\mathbf{A}$ to obtain the final spatial attention map $\mathbf{E} \in \mathbb{R}^{C\times H\times W}$:

$$E_j = \alpha \sum_{i=1}^{N} (s_{ji} D_i) + A_j, \tag{6}$$

where α is initialized as 0 and learns its weight gradually. As shown in Eq. (6), the resulting spatial attention map $\mathbf{E}$ at each position is a weighted sum of all positions and original features. Therefore, the long-range global contextual information in the spatial dimension is learned as $\mathbf{E}$.

As shown in Fig. 7, we obtain the outcome of applying each channel and spatial attention module to the input feature maps obtained through the first step. Therefore, we can establish new discriminative feature maps that consider the correlations of all pixels positions and channels in the input feature map. Later, we connected the channel attention map, spatial attention map, and input feature map according to the channel axes. The concatenated feature map is then passed through the GAP and FC layers to obtain the final 1,024-dimensional embedding vector.

D. LOSS FUNCTION

As illustrated in Fig. 3, the DNNet framework comprising two modules is optimized using two objectives, i.e., contrastive loss [31] and ArcFace loss [32]. We always consider positive or negative pairs in each DNNet input branch of the Siamese network structure for learning the robust and discriminative features. The embedding vectors acquired with the network are applied as the inputs to each loss.

Our first objective involves the contrastive loss for network optimization. The main reason for the contrastive loss is to increase the inter-class distance (negative pairs) while reducing the intra-class (positive pairs) distance. The contrastive loss can be expressed as follows:

$$\mathcal{L}_{con}(i, x_1, x_2) = (1 - i)\{max(0, m - d)\}^2 + id^2, \tag{7}$$

where x_1 is the embedding vector of the input anchor image, x_2 is the embedding vector of the corresponding positive or negative pair of the input anchor image, d is the Euclidean distance between two embedding vectors, m is the margin defining the separability in the embedding space, and i is a binary check label that distinguishes positive from negative in the pair. Here, $i = 1$ if x_1 and x_2 are positive pairs, and $i = 0$ if x_1 and x_2 are negative pairs.

Our second objective involves the ArcFace loss for network optimization. ArcFace loss maximizes the decision boundaries through margin-based representation in angular space to determine the discriminative features. The ArcFace loss can be expressed as follows:

$$\mathcal{L}_{arc} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))} + \sum_{j=1, j\neq y_i}^{n} e^{scos\theta_j}}, \tag{8}$$

where N and n are the batch size and the class number, respectively, θ_{y_i} is the target (ground truth) angle, m is the angular margin penalty, and s is the feature scale.

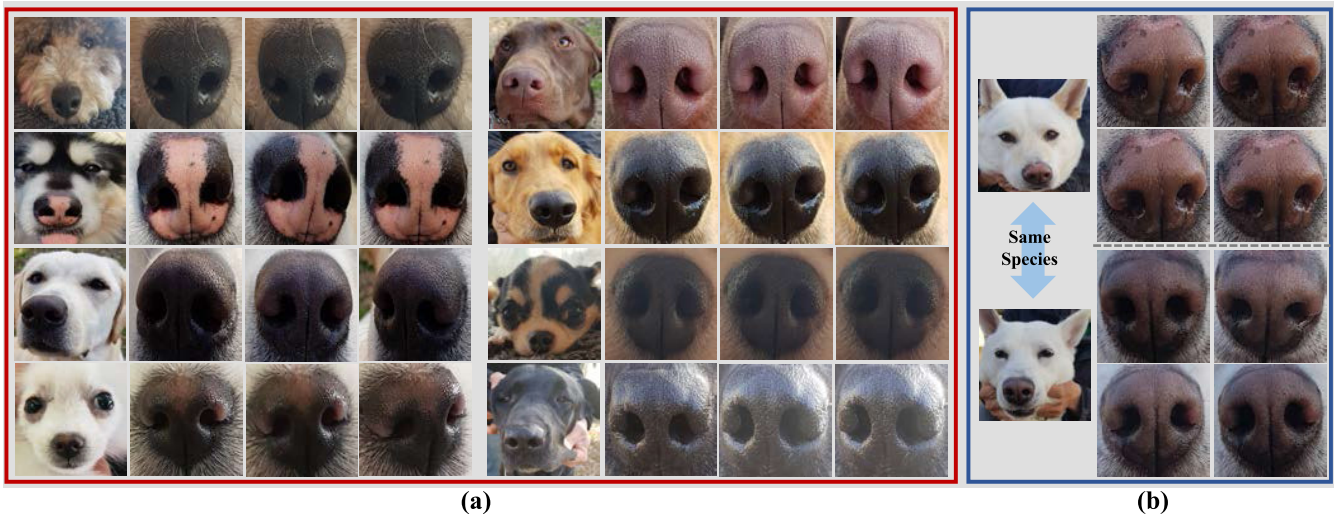

FIGURE 8. Sample images of (a) different species and (b) same species for collected dog nose-print dataset. Each class consists of dog images and corresponding nose-print images.

The embedding vector of each of the input anchor images and the corresponding positive or negative pair images are applied to the ArcFace loss.

We optimize the two modules of the proposed DNNet framework in a unified and end-to-end manner, with the full objectives being a combination of two objectives as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{con} + \frac{1}{2}\left(\mathcal{L}_{arc(anchor)} + \mathcal{L}_{arc(pair)}\right). \quad (9)$$

IV. EXPERIMENTS

A. DATASETS

In this study, we use the dog's nose-print pattern to identify individual dogs. The nose-print pattern has the only pattern that can be used as an individually identifiable means of biometric authentication, such as human fingerprints, as shown in Fig. 8. These nose-prints also have the advantage of not changing over time. Therefore, nose-prints are used to enable individual identification regardless of species. However, because it is difficult to find or obtain public datasets for the dos nose-print dataset, the dataset is obtained directly. Several shelters were visited to collect datasets. Each dog was identified with its name tag. Therefore, there are no duplicated IDs in the dataset. The dataset images were collected outside under sunlight or inside under high-intensity lamps. As illustrated in Fig. 2, the dataset was collected using mobile phones without extra scanning equipment. This dramatically increases the convenience and efficiency of data collection and processing using mobile devices. The photos were taken with a resolution of 4,032 pixels in the horizontal direction and 3,024 pixels in the vertical direction, and the nose areas were cropped manually. Only those nose-print images with more than 640 pixels were selected for inclusion in the dataset. Finally, 2,561 dog nose-print images from 302 dogs were collected for the dataset.

TABLE 1. Confusion matrix scheme.

		Predict	
		Positive	Negative
Actual	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

B. EXPERIMENTAL SETUP

1) IMPLEMENTATION DETAILS

In experiments, our networks were implemented using Pytorch [48]. The experiments were conducted on a desktop computer with an Intel(R) Core(TM) i7 CPU @ 3.20 GHz and 16.0 GB RAM.

Moreover, all the networks in this study were learned using NVIDIA RTX 2080 Ti GPU. Before performing the classification with the proposed method, the collected nose-print dataset input images were resized to 256 × 256 pixels. The batch size used was 16, and the network was trained for 200 epochs. As shown in Eq. (9), two objectives were simultaneously considered to optimize the network in an end-to-end manner. The fixed hyperparameter of contrastive loss was $m = 2$, as shown in Eq. (7). To optimize the proposed the DNNet, we used the Adam [49] optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Furthermore, as shown in Eq. (8), the hyperparameters s and m for ArcFace loss were set to 30 and 0.5, respectively. We optimized the network module responsible for ArcFace using the stochastic gradient descent (SGD) method, where the momentum was 0.9 and weight decay was 0.0005. The initial learning rate was 0.0001, which was maintained over the first 100 epochs and linearly decayed to zero over the next 100 epochs. The embedding vector size used for the feature matching was set to 1,024-dimensions.

TABLE 2. Ablation studies of proposed method on collected dog nose-print dataset, where S is siamese network structure, C is contrastive loss, A is attention module (DAN), and M is margin-based loss (ArcFace). Each backbone, with nothing selected in module and loss, is optimized using cross-entropy loss function.

Models							
Backbone	Module & Loss				Rank-1 Acc(%)	VR@FAR=0.1%(%)	VR@FAR=0.01%(%)
	S	C	A	M			
ResNet-50 [29]	-	-	-	-	93.073	45.0	36.3
	✓	✓			97.098	59.5	42.5
	✓	✓	✓		97.868	66.2	53.5
	✓	✓		✓	98.681	62.5	47.2
	✓	✓	✓	✓	98.816	71.5	57.7
ResNet-101 [29]	-	-	-	-	94.381	41.7	28.3
	✓	✓			97.154	60.5	47.1
	✓	✓	✓		96.759	59.5	43.5
	✓	✓		✓	98.518	64.1	53.2
	✓	✓	✓	✓	98.588	69.5	63.3
ResNet-152 [29]	-	-	-	-	93.956	44.2	33.3
	✓	✓			97.188	60.2	42.8
	✓	✓	✓		97.341	62.2	38.3
	✓	✓		✓	98.028	57.2	49.5
	✓	✓	✓	✓	**98.972**	**72.2**	**63.5**
VGG-19 [50]	-	-	-	-	71.751	-	-
	✓	✓			86.849	23.0	15.0
	✓	✓	✓		90.749	34.2	21.3
	✓	✓		✓	91.622	20.0	12.5
	✓	✓	✓	✓	93.230	10.8	5.8
Inception-ResNet-v2 [51]	-	-	-	-	93.126	43.8	35.8
	✓	✓			96.977	63.5	41.5
	✓	✓	✓		96.466	51.3	31.8
	✓	✓		✓	98.414	62.5	49.5
	✓	✓	✓	✓	98.624	61.8	50.5
DenseNet-121 [52]	-	-	-	-	92.229	38.3	29.2
	✓	✓			97.017	56.5	41.8
	✓	✓	✓		96.067	49.8	36.2
	✓	✓		✓	97.975	60.0	48.8
	✓	✓	✓	✓	98.776	67.8	48.2
PeleeNet [53]	-	-	-	-	92.441	38.8	26.3
	✓	✓			94.728	46.2	29.5
	✓	✓	✓		93.475	26.8	13.8
	✓	✓		✓	97.327	58.0	51.2
	✓	✓	✓	✓	97.446	35.2	32.1

2) EVALUATION METHODS

We performed ablation studies and comparisons with other state-of-the-art methods.

Furthermore, we present all experimental results through the five-fold cross-validation, a method that can be considered at this time because it is difficult to determine the generalization performance of the model only with validation results when there are insufficient datasets.

The basis of performance evaluation follows the confusion matrix shown in Table 1. The accuracy of identification is achieved through feature matching of the acquired embedded vectors for the testing set, as shown in Eq. (10). The performance was also evaluated by the receiver operating characteristic (ROC) curve and the verification rate of the specific false acceptance rate (FAR) using Eq. (11) and Eq. (13). Furthemore, the confusion matrix and the t-distributed stochastic neighbor embedding (t-SNE) [54] algorithm were used for performance evaluation.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}. \tag{10}$$

$$FAR = \frac{FP}{FP + TN}. \tag{11}$$

$$FRR = \frac{FN}{TP + FN}. \tag{12}$$

$$TAR = 1 - FRR = \frac{TP}{TP + FN}. \tag{13}$$

C. EXPERIMENTAL RESULTS

1) ABLATION STUDIES

We analyze how modules and objective functions of the proposed DNNet framework affect the system performance through the ablation studies, as shown in Table 2. We conducted the experiment by selecting the backbone network,

TABLE 3. Comparison with attention modules on collected dog nose-print dataset.

Models		Rank-1 Acc(%)	VR@FAR=0.1%(%)	VR@FAR=0.01%(%)
Proposed Method	Attention Module			
DNNet	SE [42]	92.640	21.9	-
	CBAM [43]	95.932	48.3	42.5
	SDAB [46]	96.556	23.3	16.3
	DAN [30]	**98.972**	**72.2**	**63.5**

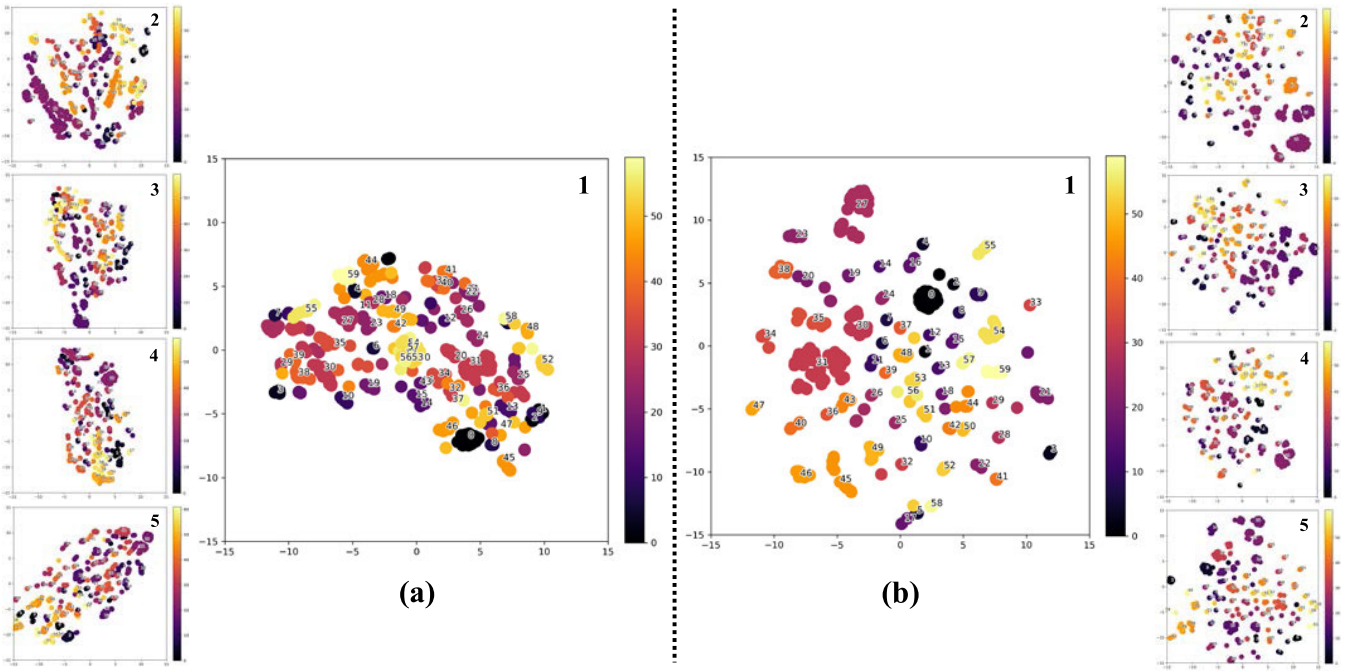

FIGURE 9. Results of t-SNE visualization of each testing set by five-fold cross-validation. Numbers in the upper-right corners correspond to multiple test sets. (a) Results of baseline model with the ResNet-152 backbone network and (b) results of the proposed DNNet framework.

the Siamese network structure, attention module, and combinations of objective functions as the factors affecting system performance. For a fair comparison, the cross-entropy loss function is used if nothing is selected other than backbone network as a baseline model. If all components, such as the Siamese network [28], contrasive loss [31], attention module (DAN) [30], and margin-based loss (ArcFace loss) [32], are selected for each backbone, it is the same as the proposed DNNet framework. As shown in Table 2, we conducted performance evaluations with the Rank-1 accuracy, VR@FAR $=$ 0.1%, and VR@FAR $=$ 0.01%.

The results of the ablation study show the lowest performance when no other option other than the backbone network is selected. On the other hand, like the DNNet method we design, we illustrate overwhelming performance across all experimental results combining the backbone network and all modules and losses. The ResNet-152 backbone with all additional options selected shows the highest performance in Rank-1 at 98.972% and the highest performance in verification tasks at VR@FAR $=$ 0.1% and VR@FAR $=$ 0.01%.

In Table 3, we present the ablation study results of DNNet performance under various attention networks. For a comparison of equivalent performance between attention modules, all conditions are trained equally except for the attention module of the DNNet framework shown in Fig. 3. We illustrate that the DNNet performance with DAN is superior to the results of applying SENet [42], CBAM [43], SDAB [46] on all performance metrics, and 2.416% higher than the second-best performance for Rank-1 accuracy.

Fig. 9 and Fig. 10 graphically depict the visualization results of the ablation studies. We present the results of each testing set using five-fold cross-validation for the baseline model and our proposed DNNet framework. First, we used the t-SNE [54] algorithm for visualization. t-SNE represents high-dimensional embedding vectors in a two-dimensional map for embedded spaces and the corresponding cluster. As shown in Fig. 9, the results of all multiple testing sets illustrate that our DNNet framework outperforms the baseline model when clustered by class, indicating that the embedding vectors' discriminative power is strong. Second, we visualized the corresponding results of the confusion matrix of

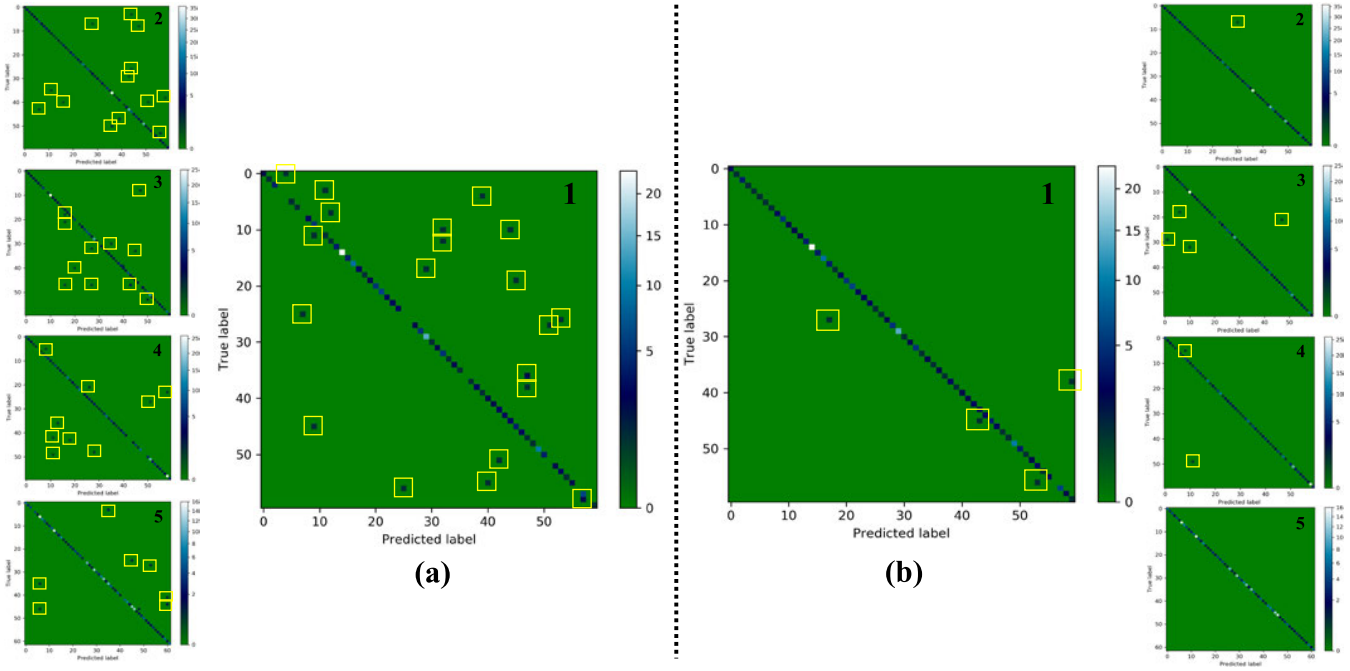

FIGURE 10. Results of confusion matrix visualization for each testing set by five-fold cross-validation. Numbers in the upper-right corners correspond to multiple test sets. (a) Results of baseline model with ResNet-152 backbone network and (b) results of proposed DNNet framework.

TABLE 4. Comparisons with other methods on collected dog nose-print dataset.

Models	Rank-1 Acc(%)	VR@FAR =0.1%(%)	VR@FAR =0.01%(%)
Hou [16]	97.081	26.3	18.8
ResNet-50 [29]	93.073	45.0	36.3
ResNet-101 [29]	94.381	41.7	28.3
ResNet-152 [29]	93.956	44.2	33.3
VGG-19 [50]	71.751	-	-
Inception-ResNet-v2 [51]	93.126	43.8	35.8
DenseNet-121 [52]	92.229	38.3	29.2
PeleeNet [53]	92.441	38.8	26.3
Taha [19]	95.718	49.2	36.3
Tharwat [27]	94.366	41.7	29.2
SIFT [55]	96.212	56.2	46.2
GLOH [56]	95.459	48.8	37.5
LBP [57]	95.502	-	-
MLBP [58]	95.778	7.6	6.6
DNNet (Ours)	**98.972**	**72.2**	**63.5**

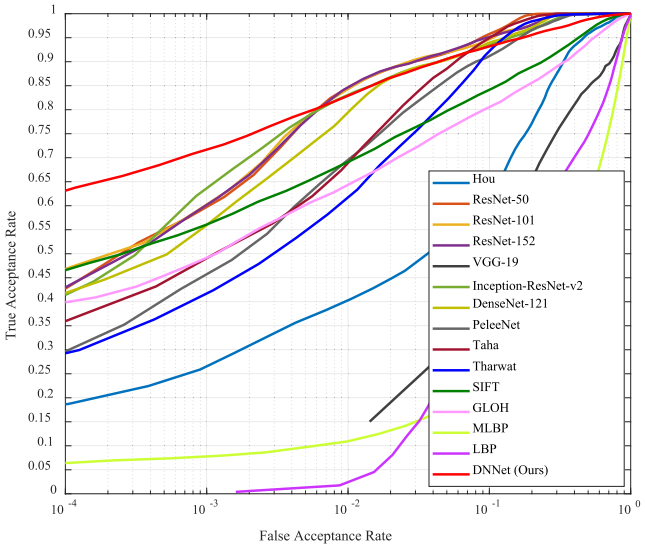

FIGURE 11. Comparisons of ROC curves of other methods on collected dog nose-print dataset.

Table 1 for the prediction label (x-axis), true label (y-axis), and color bar, where white represents the highest score. As shown in Fig. 10, the results of all multiple testing sets illustrate that the baseline model provides many incorrect predictions, whereas the DNNet framework does not. This confirms that the embedding vectors of the proposed DNNet are more discriminative than the baseline model alone.

2) COMPARISON WITH OTHER METHODS

In Table 4 and Fig. 11, we compare our DNNet framework with other handcrafted and recent deep-learning-based methods. The CNN-based deep-learning architectures compared are as follows: Hou *et al.* [16], ResNet-50/101/152 [29], SE-ResNet-50 [42], CBAM-ResNet-50 [43], Inception-ResNet-v2 [51], DenseNet-121 [52], and PeleeNet [53].

For a fair comparison for compared methods, we use the cross-entropy loss and fix the training environment. The handcrafted methods compared are as follows: Taha *et al.* [19], Tharwat *et al.* [27], SIFT [55], gradient location orientation histogram (GLOH) [56], LBP [57], and

MLBP [58]. SIFT, GLOH, LBP, and MLBP used the sliding window to deter patch overlaps when using 64×64 patches to extract the features. Therefore, we have extracted and applied 2,048, 4,352, 944, and 3,776 dimensions, respectively, for SIFT, GLOH, LBP, and MLBP. We performed performance evaluations with the Rank-1 accuracy, VR@FAR $=$ 0.1%, VR@FAR $=$ 0.01%, and ROC curve. As shown in Table 4, the proposed method of Hou *et al.* [16] illustrates the highest Rank-1 accuracy among the methods being compared. Among comparable methods, handcrafted methods outperform deep learning-based methods on average. The reason for this is that learning models through deep learning methods can be affected by the scale of the dataset. However, our DNNet method exhibits overwhelming performance for all metrics. Our DNNet method shows a Rank-1 accuracy of 98.987%, demonstrating superior performance among deep learning and handcrafted methods.

We plot the ROC curves of the proposed method and the other methods in Fig. 11. A semi-logarithmic coordinate was used to illustrate the analysis results more accurately on the ROC curve. In the ROC curves, our DNNet method is generally stable compared to the other methods. The verification rate between FAR 0.0001 and 0.006 is exceptionally effective compared to the other methods. In contrast, among the comparison methods, LBP and MLBP methods are significantly lower in performance for all FARs.

V. CONCLUSION

This paper proposes a novel dog-nose network (DNNet) deep-learning framework for individual identification of dogs using their nose-print patterns. Our method is the first attempt to identify an individual dog's nose-print patterns based on deep learning models. The DNNet method aims to obtain robust and discriminative features that can extract the unique patterns in a dog's nose prints. As ablation studies demonstrate, the performance of combining objective functions for network optimization with integrated modules that constitute DNNet is more stable than using only part of the module or a single objective function. Accordingly, the proposed DNNet enables more stable and discriminative feature extraction to identify features using the dog nose-print patterns. Moreover, our experiments demonstrate that our proposed approach outperforms state-of-the-art methods on the collected dog nose-print dataset. Consequently, the proposed DNNet method can serve as a robust baseline for individual identification. In future work, we will discuss improvements in individual identification systems by extending the nose-print dataset. We also plan to obtain a dataset for additional animals, such as cats. As previously noted in related studies, nose-print patterns are important feature extractions that distinguish species characteristics. Therefore, we will apply it to the task of identifying animal species.

ACKNOWLEDGMENT

Petnow Inc. supported data acquisition, organization, and validation of the experiments.

REFERENCES

[1] H. S. Kühl and T. Burghardt, "Animal biometrics: Quantifying and detecting phenotypic appearance," *Trends Ecol. Evol.*, vol. 28, no. 7, pp. 432–441, Jul. 2013.
[2] J. Duyck, C. Finn, A. Hutcheon, P. Vera, J. Salas, and S. Ravela, "Sloop: A pattern retrieval engine for individual animal identification," *Pattern Recognit.*, vol. 48, no. 4, pp. 1059–1073, Apr. 2015.
[3] S. Kumar, A. Pandey, K. S. R. Satwik, S. Kumar, S. K. Singh, A. K. Singh, and A. Mohan, "Deep learning framework for recognition of cattle using muzzle point image pattern," *Measurement*, vol. 116, pp. 1–17, Feb. 2018.
[4] S. Kumar, S. K. Singh, R. S. Singh, A. K. Singh, and S. Tiwari, "Real-time recognition of cattle using animal biometrics," *J. Real-Time Image Process.*, vol. 13, no. 3, pp. 505–526, Sep. 2017.
[5] S. Kumar, S. Tiwari, and S. K. Singh, "Face recognition for cattle," in *Proc. 3rd Int. Conf. Image Inf. Process. (ICIIP)*, Dec. 2015, pp. 65–72.
[6] S. Kumar, S. Tiwari, and S. K. Singh, "Face recognition of cattle: Can it be done?" *Proc. Nat. Acad. Sci. USA*, vol. 86, no. 2, pp. 137–148, Jun. 2016.
[7] H. Wang, H. Su, P. Chen, R. Hou, Z. Zhang, and W. Xie, "Learning deep features for giant panda gender classification using face images," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 279–285.
[8] G. Mougeot, D. Li, and S. Jia, "A deep learning approach for dog face verification and recognition," in *Proc. Pacific Rim Int. Conf. Artif. Intell.* Springer, 2019, pp. 418–430.
[9] R. Kumar, M. Sharma, K. Dhawale, and G. Singal, "Identification of dog breeds using deep learning," in *Proc. IEEE 9th Int. Conf. Adv. Comput. (IACC)*, Dec. 2019, pp. 193–198.
[10] W. M. Matkowski, A. W. K. Kong, H. Su, P. Chen, R. Hou, and Z. Zhang, "Giant panda face recognition using small dataset," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1680–1684.
[11] S. Taheri and Ö. Toygar, "Animal classification using facial images with score-level fusion," *IET Comput. Vis.*, vol. 12, no. 5, pp. 679–685, Aug. 2018.
[12] Q. He, Q. Zhao, N. Liu, P. Chen, Z. Zhang, and R. Hou, "Distinguishing individual red pandas from their faces," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*. Springer, 2019, pp. 714–724.
[13] M. F. Hansen, M. L. Smith, L. N. Smith, M. G. Salter, E. M. Baxter, M. Farish, and B. Grieve, "Towards on-farm pig face recognition using convolutional neural networks," *Comput. Ind.*, vol. 98, pp. 145–152, Jun. 2018.
[14] D. Deb, S. Wiper, S. Gong, Y. Shi, C. Tymoszek, A. Fletcher, and A. K. Jain, "Face recognition: Primates in the wild," in *Proc. IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Oct. 2018, pp. 1–10.
[15] D. Crouse, R. L. Jacobs, Z. Richardson, S. Klum, A. Jain, A. L. Baden, and S. R. Tecot, "LemurFaceID: A face recognition system to facilitate individual identification of lemurs," *BMC Zool.*, vol. 2, no. 1, p. 2, Dec. 2017.
[16] J. Hou, Y. He, H. Yang, T. Connor, J. Gao, Y. Wang, Y. Zeng, J. Zhang, J. Huang, B. Zheng, and S. Zhou, "Identification of animal individuals using deep learning: A case study of giant panda," *Biol. Conservation*, vol. 242, Feb. 2020, Art. no. 108414.
[17] N. Coldea, "Nose prints as a method of identification in dogs," *Veterinary Quart.*, vol. 16, no. 1, p. 60, 1994.
[18] A. I. Awad, A. E. Hassanien, and H. M. Zawbaa, "A cattle identification approach using live captured muzzle print images," in *Proc. Int. Conf. Secur. Inf. Commun. Netw.* Springer, 2013, pp. 143–152.
[19] A. Taha, A. Darwish, and A. E. Hassanien, "Arabian horse identification system based on live captured muzzle print images," in *Proc. Int. Conf. Adv. Intell. Syst. Inform.* Springer, 2017, pp. 778–787.
[20] W. Kusakunniran, A. Wiratsudakul, U. Chuachan, S. Kanchanapreechakorn, and T. Imaromkul, "Automatic cattle identification based on fusion of texture features extracted from muzzle images," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Feb. 2018, pp. 1484–1489.
[21] A. I. Awad and M. Hassaballah, "Bag-of-visual-words for cattle identification from muzzle print images," *Appl. Sci.*, vol. 9, no. 22, p. 4914, Nov. 2019.
[22] A. Noviyanto and A. M. Arymurthy, "Beef cattle identification based on muzzle pattern using a matching refinement technique in the SIFT method," *Comput. Electron. Agricult.*, vol. 99, pp. 77–84, Nov. 2013.
[23] S. Kumar, S. K. Singh, A. I. Abidi, D. Datta, and A. K. Sangaiah, "Group sparse representation approach for recognition of cattle on muzzle point images," *Int. J. Parallel Program.*, vol. 46, no. 5, pp. 812–837, Oct. 2018.
[24] Y.-C. Chen, S. C. Hidayati, W.-H. Cheng, M.-C. Hu, and K.-L. Hua, "Locality constrained sparse representation for cat recognition," in *Proc. Int. Conf. Multimedia Modeling*. Springer, 2016, pp. 140–151.

[25] S. Chakraborty, K. Karthik, and S. Banik, "Investigation on the muzzle of a pig as a biometric for breed identification," in *Proc. 3rd Int. Conf. Comput. Vis. Image Process.* Springer, 2020, pp. 71–83.
[26] A. Noviyanto and A. M. Arymurthy, "Automatic cattle identification based on muzzle photo using speed-up robust features approach," in *Proc. 3rd Eur. Conf. Comput. Sci. (ECCS)*, vol. 110, 2012, p. 114.
[27] A. Tharwat, T. Gaber, A. E. Hassanien, H. A. Hassanien, and M. F. Tolba, "Cattle identification using muzzle print images based on texture features approach," in *Proc. 5th Int. Conf. Innov. Bio-Inspired Comput. Appl. (IBICA)*. Springer, 2014, pp. 217–227.
[28] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, Lille, France, vol. 2, 2015.
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[30] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[31] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1735–1742.
[32] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.
[33] T. Chehrsimin, T. Eerola, M. Koivuniemi, M. Auttila, R. Levänen, M. Niemi, M. Kunnasranta, and H. Kälviäinen, "Automatic individual identification of Saimaa ringed seals," *IET Comput. Vis.*, vol. 12, no. 2, pp. 146–152, Mar. 2018.
[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[35] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[36] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[37] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.
[38] M. Favorskaya and A. Pakhirka, "Animal species recognition in the wildlife based on muzzle and shape features using joint CNN," *Procedia Comput. Sci.*, vol. 159, pp. 933–942, Jan. 2019.
[39] H. Hu, B. Dai, W. Shen, X. Wei, J. Sun, R. Li, and Y. Zhang, "Cow identification based on fusion of deep parts features," *Biosystems Eng.*, vol. 192, pp. 245–256, Apr. 2020.
[40] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167
[41] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*, 2010, pp. 807–814.
[42] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[43] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[44] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[45] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "A^2-nets: Double attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 352–361.
[46] C.-Y. Lin, Z. Tao, A.-S. Xu, L.-W. Kang, and F. Akhyar, "Sequential dual attention network for rain streak removal in a single image," *IEEE Trans. Image Process.*, vol. 29, pp. 9250–9265, 2020.
[47] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2019, pp. 7354–7363.
[48] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. NIPS Autodiff Workshop Future Gradient-Based Mach. Learn. Softw. Techn.*, 2017.
[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[51] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," 2016, *arXiv:1602.07261*. [Online]. Available: http://arxiv.org/abs/1602.07261
[52] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[53] R. J. Wang, X. Li, and C. X. Ling, "PELEE: A real-time object detection system on mobile devices," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1963–1972.
[54] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[55] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[56] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.
[57] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[58] B. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 639–646, Mar. 2011.

HAN BYEOL BAE received the B.S. in electrical and electronic engineering from Yonsei University, Seoul, in 2010, the B.S. degree in information and communication engineering from Yonsei University, Wonju, in 2010, and the M.S. degree in biometrics engineering and the Ph.D. degree in electrical and electronic engineering from Yonsei University, in 2015 and 2020, respectively. He is currently a Postdoctoral Researcher with the Image and Video Pattern Recognition Laboratory, Yonsei University. His research interests include face recognition, image translation, and image classification.

DAEHYUN PAK received the bachelor's degree from Yonsei University, in 2009, and the Ph.D. degree in electrical and electronic engineering from the Graduate School of Yonsei University, in 2019. His research interests include a real-time deep-neural-network application of image/video recognition system on hand-held devices, and general applications of DNN-based video signal processing.

SANGYOUN LEE (Member, IEEE) received the B.S. and M.S. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 1987 and 1989, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 1999. He is currently a Professor and the Head of Electrical and Electronic Engineering with the Graduate School, and the Head of the Image and Video Pattern Recognition Laboratory, Yonsei University. His research interests include all aspects of computer vision, with a special focus on pattern recognition for face detection and recognition, advanced driver-assistance systems, and video codecs.