Combining Deep Convolutional Neural Networks With Stochastic Ensemble Weight Optimization for Facial Expression Recognition in the Wild

Jae Young Choi, *Member, IEEE*, and Bumshik Lee, *Member, IEEE*

***Abstract*—Although recent emotion recognition methods (based on facial expression cues) achieve excellent performance in controlled scenarios, the recognition of emotion in the wild remains a challenging problem because of occlusion, large head poses, illumination variations, etc. Recent advances in deep learning show that combining an ensemble of deep learning models can considerably outperform the approach of using only a single deep learning model for challenging recognition problems. This paper presents a novel ensemble deep learning method, "deep convolutional neural network (DCNN) ensemble classifier", for improved facial expression recognition (FER) in the wild. Our proposed DCNN ensemble classifier is novel in terms of the following aspects: (1) the process of finding ensemble weights for combining DCNN decision outputs is formulated as a stochastic optimization problem (via simulated annealing) in which the energy to be minimized represents the generalized (test) classification error of the DCNN ensemble and (2) for the creation of DCNN ensemble members, we propose the combined use of different types of face representations and bagging (T. G. Dietterich, 2000), which is quite useful in increasing the diversity of the DCNN ensemble. Extensive and comparative experiments on three wild FER datasets, namely FER2013, SFEW2.0, and RAF-DB, show that the proposed DCNN ensemble classifier achieves competitive FER performances when compared with other recently developed methods—76.69%, 58.68%, and 87.13% of FER accuracy under the FER2013, SFEW2.0, and RAF-DB evaluation protocols, respectively.**

***Index Terms*—Facial expression recognition, ensemble deep learning, stochastic ensemble weight optimization, simulated annealing, energy function, deep ensemble generalization error.**

Manuscript received 30 January 2021; revised 31 May 2021 and 13 August 2021; accepted 3 October 2021. Date of publication 26 October 2021; date of current version 13 January 2023. This work was supported in part by the Hankuk University of Foreign Studies Research Fund, in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) under Grant 2021R1A2C1092322, in part by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) under Grant 1711134404. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Ramanathan Subramanian. *(Corresponding author: Bumshik Lee.)*

Jae Young Choi is with the Division of Computer Engineering, Hankuk University of Foreign Studies, Yongin-si, Gyeonggi-do 17025, Korea (e-mail: jychoi@hufs.ac.kr).

Bumshik Lee is with the Information and Communication Engineering, Chosun University, Gwangju 61452, Korea (e-mail: bslee@chosun.ac.kr).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2021.3121547.

Digital Object Identifier 10.1109/TMM.2021.3121547

I. Introduction

Facial expression traits play the most important role in analyzing a subject's emotional state in a spontaneous environment [18]. Automated facial expression recognition (FER) has a wide range of applications, such as in interactive games, sociable robotics, crowd analytics, neuromarketing, and several other human-computer interaction systems [49]. Recently, with the transition of FER from laboratory-controlled to challenging in-the-wild conditions, deep convolutional neural networks (DCNNs) [12], [20] have achieved state-of-the-art performances across a variety of FER-related applications [50]. Thus, the DCNNs have been mainstreamed in the field of FER; the results so far have been promising, and most of the FER challenge winners have employed DCNNs [3]–[5], [15]. It mainly benefits from large-scale training data and an end-to-end learning framework [20].

Thus far, almost all DCNN-based FER systems are based on the *design paradigm of a single classifier* with only one input image representation (grayscale or *RGB* color representations) [54]. However, the FER problem in the wild is too complex [21], [22] to be solved by a single DCNN-based classifier. This is because facial images are often captured in the wild under natural conditions such as occlusion, large head pose or illumination variations, and lower image resolution [21], [22]. As such, it is difficult to separate the expressions' feature space, that is, facial features from one subject in two different expressions may be quite close in the feature space, while the facial features from two subjects with the same expression may be very far from each other [51], [52].

Many examples of both natural and artificial systems show that a composite system consisting of several subsystems can reduce the total complexity of the system while satisfactorily solving a difficult problem [23], [26]. The success of a neural network ensemble in improving a classifier's generalization [24], [25] is a typical example. In pattern classification, ensemble classifiers are often considered a practical and effective solution for difficult recognition problems such as in-the-wild FER [28]. The use of ensemble classifiers is supported by the fact that the patterns misclassified by different classifiers are not necessarily the same [14], [26]. However, designing an ensemble of multiple DCNNs is still an open problem [27]. In particular, effectively constructing an ensemble of DCNNs and combining their

1520-9210 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: HANKUK UNIVERSITY OF FOREIGN STUDIES. Downloaded on January 17,2023 at 18:31:11 UTC from IEEE Xplore. Restrictions apply.

decisions can be the primary challenge and an important research direction [27], [54].

In this paper, we propose a new DCNN ensemble classifier for in-the-wild FER. Our method has two key aspects. First, we formulate the process of finding ensemble combination weights as an optimization problem, which can be effectively solved using our novel *simulated annealing (SA)-based algorithm*. We show that ensemble weights optimized via our SA-based algorithm can significantly improve FER performance. Second, we introduce an effective DCNN ensemble construction that takes advantage of the combined use of different face representations and bagging [16]. This allows maximizing the diversity among DCNN ensemble members in the sense that they do not make coincident errors, as demonstrated by our experimental results in Section VI.C. We evaluate the proposed method on three publicly available wild FER databases (DBs) collected under real-world scenarios: FER2013 [29], static facial expressions in the wild (SFEW2.0) [53], and real-world affective face database (RAF-DB) [67]. The results of our method are better than, or at least competitive with the best-reported results.

The rest of this paper is organized as follows: Section II reviews previous works on deep learning-based FER. Section III describes the motivation behind the proposed DCNN ensemble weight optimization and construction for completeness. Section IV details the proposed DCNN ensemble construction. In Section V, the proposed SA-based DCNN ensemble weight optimization algorithm is presented. Section VI presents extensive and comparative experimental results to demonstrate the effectiveness of the proposed method for FER in the wild. Finally, the discussion and conclusions are presented in Section VII.

II. Related Work

A significant part of FER's recent progress has been achieved because of the emergence of deep learning models and, more specifically, with DCNNs [12]. In the following paragraphs, we review some recent methods based on deep neural networks that are most relevant to our work and refer the reader to a recent and comprehensive survey [54] on FER using deep learning for further information.

The authors of [55] proposed an action unit-inspired deep network (AUDN) to exploit a psychological theory where expressions can be decomposed into several facial expression action units. Tang [46] proposed a deep neural network architecture using a linear support vector machine (SVM) as the top layer instead of a softmax layer. To train deep neural networks, L2-SVM loss was used for facial expression classification. Devries *et al.* [47] developed a multi-task CNN that jointly predicts FER and facial landmarks and demonstrated that learning features associated with facial landmark position can improve FER. Shao *et al.* [19] proposed three different DCNN models, namely shallow Light-CNN, dual-branch CNN, and pre-trained CNN, for FER in the wild. Liu *et al.* [56] proposed a method for optimizing both (N+M)-tuples cluster loss and softmax loss via two fully connected layer branch configurations. Specifically, during the training, the (N+M)-tuples cluster loss was formalized to alleviate the difficulty of anchor selection and threshold validation in the triplet loss for the identity-invariant FER. Ding *et al.* [57] proposed the FaceNet2ExpNet, where a novel two-stage training algorithm for FER was used. In the first stage, a probabilistic distribution function was used to regularize the training of the targeted FER net based on the already fine-tuned face net. In the second stage, to further boost the discriminative capability, randomly initialized fully connected layers attached to the trained convolutional blocks were used to train the entire network from scratch with strong facial expression label supervision. Meng *et al.* [58] proposed a novel identity-aware CNN (IACNN) that contains two identical sub-CNNs with sharing weights. In their method, the expression-sensitive contrastive loss was used to learn expression-related features, while the identity-sensitive contrastive loss was employed to learn identity-related features. Identity-invariant FER was achieved by combining expression-related and identity-related features.

Recent advances in deep learning suggest that *the use of an ensemble of DCNNs can improve the performance of image recognition and classification tasks*. In [1], multi-column deep neural networks have been suggested, where each column is represented as a DCNN. The outputs of all columns are simply averaged for decision aggregation. Yu *et al.* [4] proposed an effective optimization framework using the log-likelihood loss and hinge loss to adaptively combine multiple DCNNs to perform expression recognition. Kim *et al.* [5] used an exponentially weighted fusion based on validation accuracy and constructed a hierarchical architecture of DCNN committees by implementing majority voting or a simple average for FER. Pramerdorfer *et al.* [59] constructed an ensemble of DCNNs based on different deep architectures such as VGG, Inception, and ResNet; they used a simple ensemble voting of the outputs produced by eight DCNNs as ensemble members. In summary, the aforementioned works show that deep learning models can be employed as good base classifiers (i.e., ensemble members) for typical ensemble classification approaches, such as the average combination rule [28]. *However, ensemble combination approaches with deep learning in the field of FER [54] have been limited to simple averaging or majority voting based on the conditional probability vector [28] obtained from each DCNN*. Thus, **an optimal combination of the outputs of ensemble DCNN classifiers has not yet been explored in the literature**.

Unlike other recent DCNN-based ensemble approaches focused on FER [1], [4], [5], [7], [54], [59], the main contributions of our work are summarized as follows:

- The focus of this work is on **making the DCNN ensemble combination more adaptable** to the recognition problem being addressed. We developed a novel method that optimally determines the ensemble (combination) weights by minimizing the generalized (test) classification error of the whole DCNN ensemble. In our method, finding ensemble weights is formulated as an **optimization problem**, and it can be effectively solved using **our novel simulated annealing (SA) [36], [37]-based algorithm** well-suited for improving FER in the wild.
- In addition to ensemble weight optimization, we propose a new ensemble construction approach that takes advantage of **the combined use of different face representations**

Authorized licensed use limited to: HANKUK UNIVERSITY OF FOREIGN STUDIES. Downloaded on January 17,2023 at 18:31:11 UTC from IEEE Xplore. Restrictions apply.

and bagging [16] to create DCNN ensemble members. It allows for maximizing diversity (for facilitating a complementary effect) among DCNN members, thereby boosting FER performance.

III. Motivation

To combine the outputs of DCNN members, we adopt weighted majority voting (WMV) because WMV and its variants are by far the most popular approach [43] for ensemble combination. Assuming that we are given a set of *M* individual DCNNs, $\{f_1, \ldots, f_M\}$, our task is to combine the outputs of $f_k (k = 1, \ldots, M)$ to predict the emotion class label c_j (e.g., happy). Here, the outputs of f_k are in the form of an *L*-dimensional vector $[f_k^1(\boldsymbol{x}), f_k^2(\boldsymbol{x}), \ldots, f_k^L(\boldsymbol{x})]$ for the input instance (sample) $\boldsymbol{x}$, where $f_k^j(\boldsymbol{x}) \in [0, 1]$ represents an estimate of the posterior probability $P(c_j|\boldsymbol{x})$. In WMV, the output of the DCNN ensemble classifier can be defined as

$$\begin{aligned} F(\boldsymbol{x}) &= \textstyle\sum_{k=1}^{M} w_k \, f_k(\boldsymbol{x}) \\ &= \textstyle\sum_{k=1}^{M} [w_k \, f_k^1(\boldsymbol{x}), w_k f_k^2(\boldsymbol{x}), \ldots, w_k f_k^L(\boldsymbol{x})] \end{aligned} \tag{1}$$

where w_k is the weight assigned to each DCNN member. In practice, the weights are usually normalized and constrained by $w_k = 0$ and $\sum_{k=1}^{M} w_k = 1$.

With adequate weight assignments, the classification performance of WMV can be significantly improved [26], [30]. In particular, Zhou's work [31] provides insight and motivation into the importance of developing an ensemble weight optimization algorithm. He adopted a Bayesian optimal discriminant analysis to determine ensemble weights for achieving a minimum ensemble classification error. We now briefly review this analysis to adequately explain our proposed method. Let p_k denote the classification accuracy of f_k and let us assume that the outputs of the individual $f_k (k = 1, \ldots, M)$ are conditionally independent (i.e., uncorrelated). Then, a Bayesian optimal discriminant function that leads to a minimum classification error for the combined output on the class label c_j can be written as [31]

$$\log P(c_j) + \sum_{k=1}^{M} f_k^j(\boldsymbol{x}) \log \frac{p_k}{1 - p_k} \tag{2}$$

where $P(c_j)$ is the prior probability of the input $\boldsymbol{x}$ being from the emotion class c_j. The second term in (2) suggests that the ensemble weights should be proportional to the classification accuracies of individual DCNNs, that is, $w_k \approx \log \frac{p_k}{1-p_k}$. This also theoretically supports the argument that different weights should be properly used for individual DCNNs relying on different strengths.

However, the weights chosen using (2) are obtained by assuming independence among the outputs of the individual classifiers. In practical applications, this does not hold because individual classifiers are usually correlated to a large extent. Hence, for the WMV combination to be effective on the DCNN ensemble, *weight optimization should be performed by considering an appropriate compromise between the DCNN classification accuracy and their correlation* (i.e., diversity [32]). To achieve this, we design a **novel stochastic optimization algorithm that finds the optimal weights** by minimizing the ensemble generalization error composed of ensemble accuracy and diversity. Another important issue is that *the individual DCNNs must be as diverse* (*uncorrelated*) *as possible* while being accurate. In light of this fact, Siqueira *et al.* [7] proposed the so-called Ensemble with Shared Representations (ESRs) based on deep convolutional networks for FER, aiming to achieve high diversity (low redundancy) in the ensemble. This motivates us to develop an **effective ensemble construction of diverse DCNNs**, which will be explained in the next section.

IV. Proposed DCNN Ensemble Construction

To construct an ensemble of DCNNs that are diverse while being accurate, we propose the combined use of different face representations and bagging. For different face representations, facial texture images such as local binary pattern (LBP) [33] or Gabor [33] are used because they have been widely used for face representations in FER [12], [54].

From the entire training set $\mathbf{T}$ containing N samples, we create P bootstrap training sets, denoted as $\mathbf{T}_p$, $p = 1, \ldots, P$, by forming bootstrap replicates of the original training samples in [16]. Therefore, each $\mathbf{T}_p$ will contain, on average, $0.6N$ different samples [16], some of them repeated once or more. The remaining $0.4N$ examples in $\bar{\mathbf{T}}_p = \mathbf{T} - \mathbf{T}_p$ are used for validation purposes in the training phase for each DCNN, that is, tuning the learning rate and determining when to stop network training [41]. Note that different face representations [33] are obtained by transforming an original *RGB* (or grayscale) image into various facial texture images based on associated parameters, for example, the number of sampling points and radius value for LBP face representations, as shown in Fig. 1. Each of the facial images contained in a particular $\mathbf{T}_p$ is transformed into Q different face representations, resulting in a total of PQ bootstrap training sets. These multiple transformed bootstrap training sets, each with a specific face representation, are applied as inputs to train individual DCNNs as ensemble members. This way, we generate a total of M DCNN ensemble members, where $M = PQ$.

Fig. 1 shows the proposed construction of DCNN ensemble members. To our knowledge, we are the first to propose **the combined use of various facial texture representations and bagging** [16] to learn DCNNs as ensemble members for FER. As demonstrated in our experiment (see Table III), our proposed combination of facial texture representation with bagging is advantageous for increasing diversity (i.e., decreasing correlation) among DCNN members, which is beyond the case of using only *RGB* (or grayscale) as the input space for learning DCNNs. This makes the DCNN members so different that they contradict each other, thereby boosting ensemble classification performance.

V. DCNN Ensemble Weight Optimization With Simulated Annealing

There are two main theoretical insights [32] behind the soundness of using ensemble classifier models: (a) *ambiguity decomposition* and (b) *bias-variance decomposition*, both of which

Authorized licensed use limited to: HANKUK UNIVERSITY OF FOREIGN STUDIES. Downloaded on January 17,2023 at 18:31:11 UTC from IEEE Xplore. Restrictions apply.

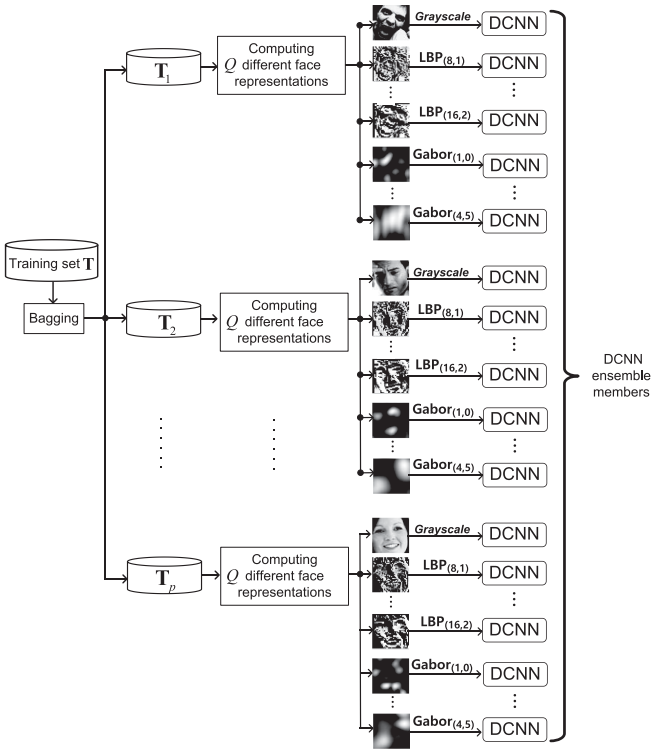


Fig. 1. Visualization for the proposed DCNN ensemble construction. The subscript enclosed in brackets indicates the following format [33]: (*No. of sampling points*, *radius*) of a circular neighborhood for LBP and (*scale*, *orientation*) of the Gabor kernel.

offer a theoretical justification for improving the generalization performances of an ensemble classifier over its base (component) classifiers. Ambiguity decomposition states that the quadratic error of the ensemble classifier is guaranteed to be less than or equal to the average quadratic error of the base classifiers. In contrast, bias-variance decomposition [9] states that the generalization error of an ensemble classifier can be broken down into two components: bias (the average difference between the prediction of the base classifier and the target output) and variance (the average variability of the base classifiers). Brown *et al.* [10], [11] showed the equivalence between ambiguity and bias-variance decompositions, and hence, there exists a common term that quantifies the **accuracy-diversity trade-off** for achieving a lower ensemble generalization error. This represents a well-grounded theoretical basis for pursuing the *right balance between accuracy and diversity* of the DCNN ensemble in the process of ensemble weight optimization. In the proposed ensemble weight optimization using SA, a measure of the balance between DCNN ensemble accuracy and diversity is given by the **energy function**, which will be explained in the next subsection.

A. Proposed Energy Function

To find ensemble weights that achieve the lowest generalized classification error, we propose the following energy function as the learning objective:

$$E_{\boldsymbol{w}}(\boldsymbol{x}) = \mathrm{Acc}_{\boldsymbol{w}}(\boldsymbol{x}) - \lambda \cdot \mathrm{Div}_{\boldsymbol{w}}(\boldsymbol{x})$$

$$= \left\| \sum_k w_k f_k(\boldsymbol{x}) - t(\boldsymbol{x}) \right\|_p - \lambda \sum_k w_k \sqrt{\frac{1}{M} \sum_{k \neq n} \| f_k(\boldsymbol{x}) - f_n(\boldsymbol{x}) \|_p} \quad (3)$$

where $f_k(\boldsymbol{x})$ outputs the estimated class probabilities on sample $\boldsymbol{x}$, $t(\boldsymbol{x})$ is the true distribution, where the entire probability mass is on a correct emotion class (i.e., if the true class of $\boldsymbol{x}$ is "angry" and is placed at the ℓ-th position, $t = [0, \ldots, 1, \ldots, 0]$ contains a single 1 at the ℓ-th position), and $\| \cdot \|_p$ denotes the p-norm of a vector; L^2-norm was used (i.e., $p = 2$) in our experiments. In (3), the first term is the **ensemble accuracy** that computes the classification error of the DCNN ensemble, while the second one characterizes **ensemble diversity**, which is a measure of how individual DCNNs' outputs (on data sample) differ from each other so that it can quantify the variability (i.e., disagreement) among DCNN members. The $\mathrm{Div}_{\boldsymbol{w}}(\boldsymbol{x})$ is based on root quartic negative correlation (RTQRT-NCL) [34].

The larger the diversity term is, the larger the ensemble generalization error reduction will be [25], [35]. However, as the variability of the individual DCNNs rises, so does the value of the first accuracy term [24]. Therefore, we need to optimize the balance between the two terms to find optimal ensemble weights. To this end, a λ parameter is introduced to control the trade-off between the two terms in (3); for $\lambda = 0$, we optimize the ensemble weight w_k for each of the DCNN members by considering only their ensemble accuracy. In contrast, as λ increases, higher weights would be imposed on DCNN members with a larger diversity. A good compromise is found by setting $\lambda = 0.3$ (for more details, see Section VI.A). Also note that a proper use of $\mathrm{Div}_{\boldsymbol{w}}(\boldsymbol{x})$ is helpful in relieving overfitting to the validation set for finding ensemble weights. The experimental results demonstrating this advantage are presented in Fig. 6.

Note that energy function $\boldsymbol{E}_{\boldsymbol{w}}(\cdot)$ is parameterized by the set of ensemble weights $\boldsymbol{w} = \{w_k : k = 1, \ldots, M\}$ for all M DCNNs. As such, minimizing $\boldsymbol{E}_{\boldsymbol{w}}(\cdot)$ with respect to $\boldsymbol{w}$ is the objective of the proposed ensemble weight optimization. Considering this, we solve the following optimization problem:

$$\boldsymbol{w}^{\mathrm{opt}} = \arg\min_{\boldsymbol{w}} E_{\boldsymbol{w}}(\boldsymbol{x}|\boldsymbol{V}) = \arg\min_{\boldsymbol{w}} \frac{1}{|\boldsymbol{V}|} \sum_{(\boldsymbol{x},t) \in \boldsymbol{V}} E_{\boldsymbol{w}}(\boldsymbol{x}) \quad (4)$$

where $\boldsymbol{V}$ denotes a validation set (of the entire dataset) that should be kept unseen by all DCNNs during their ensemble construction explained in Section III, and $|\cdot|$ is the cardinality of a set.

B. Proposed Simulated Annealing Optimization Algorithm

The optimization problem in (4) has a quadratic form in terms of multivariables $w_1, w_2, \ldots, w_M$ as $\mathrm{the}\, p - \mathrm{norm}$ is used in (3). Hence, finding $\boldsymbol{w}^{\mathrm{opt}}$ is a nonlinear multivariate optimization problem where multiple local minima may exist. In this context, gradient-based methods are likely to get stuck in unacceptable local minima [36]. To tackle this problem, we consider

Authorized licensed use limited to: HANKUK UNIVERSITY OF FOREIGN STUDIES. Downloaded on January 17,2023 at 18:31:11 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: Simulated annealing (SA)-based algorithm for finding optimal DCNN ensemble weights.

1: Input: ensemble weights $\boldsymbol{w}_0 = [w_{0,1}, \ldots, w_{0,k}, \ldots, w_{0,M}]$ and validation set $\boldsymbol{V}$.
2: Initialize ensemble weights $w_{0,k} = 1/M$ for all *k*.
3: Initialize annealing temperature $T_0 = 3E_{\boldsymbol{w}_0}(\boldsymbol{x}|\boldsymbol{V})$ using Eq. (3) and (4).
4: Initialize maximum number of annealing steps b_{max}.
5: Initialize final temperature and energy T_{final} and E_{final} at which to stop.
6: Initialize maximum number of tries within one temperature try_{max}.
7: while $b < b_{max}$ or $T_b < T_{\text{final}}$ or $E_{\boldsymbol{w}_b}(\boldsymbol{x}|\boldsymbol{V}) < E_{\text{final}}$ **do**
8: $b = b + 1$.
9: Repeat iterations:
10: Randomly select a certain weight $w_{b-1,k}$ from all *M* previous weights.
11: Perturb selected $w_{b-1,k}$ using (5), leading to $w_{b-1,k} \rightarrow w_{b-1,k} + \Delta w_k$.
12: Generate candidate weight solution $\boldsymbol{w}_b$ using Eq. (6).
13: Evaluate energy $E_{\boldsymbol{w}_b}(\boldsymbol{x}|\boldsymbol{V})$ using Eq. (3) and (4).
14: **if** $(E_{\boldsymbol{w}_b}(\boldsymbol{x}|\boldsymbol{V}) - E_{\boldsymbol{w}_{b-1}}(\boldsymbol{x}|\boldsymbol{V})) < 0$.
15: **then** $\boldsymbol{w}_b$ is the new current weight solution.
16: else if $e^{[-(E_{\boldsymbol{w}_b}(\boldsymbol{x}|\boldsymbol{V}) - E_{\boldsymbol{w}_{b-1}}(\boldsymbol{x}|\boldsymbol{V}))/T_b]} > \text{Rand}[0, 1)$.
17: then $\boldsymbol{w}_b$ is the new current weight solution.
18: until try_{max} is reached (all weight values polled several times).
19: Reduce temperature $T_b = 0.9T_{b-1}$.
20: end while
21: Output: optimal DCNN ensemble weights $\boldsymbol{w}_{\text{opt}}$.

a stochastic optimization algorithm. SA has proven successful in extremely complicated optimization tasks [37] because it can deal with local optima without getting stuck in them while searching for the global optimum.

In the following paragraphs, we describe the proposed SA algorithm for finding the best set of ensemble weights. Assuming that the initial weights are given by $\boldsymbol{w}_0 = [w_{0,1}, \ldots, w_{0,k}, \ldots, w_{0,M}]$ (e.g., $w_{0,k} = 1/M$ for all *k*), we denote the *k*-th ensemble weight at the *b*-th SA iteration by $w_{b,k}$. To generate the candidate weight solution $\boldsymbol{w}_b$ at the *b*-th iteration, one of the *M* weight variables $w_{b-1,k}(k = 1, \ldots, M)$ at the previous iteration is randomly selected and then perturbed in the following way:

$$w_{b-1,k} \rightarrow w_{b-1,k} + \Delta w_k \text{ and } \Delta\, w_k = rw_{max} \tag{5}$$

where w_{max} is the maximum value of weights (we set $w_{max} = 1$) and *r* is a random number in the interval $[-1, 1]$. Note that Δw_k represents the state change of a randomly selected weight $w_{b-1,\,k}$. Using (5), the candidate weight solution $\boldsymbol{w}_b$ is obtained as

$$\boldsymbol{w}_b = \boldsymbol{w}_{b-1} + \Delta\boldsymbol{w} \text{ and } \Delta\boldsymbol{w} = [0, \ldots, \Delta w_k, \ldots, 0] \tag{6}$$

where $\Delta\boldsymbol{w}$ contains a single Δw_k at the *k*-th position and zero otherwise. The proposed SA algorithm checks whether the energy function in (3) increases or decreases when ensemble DCNNs are combined using the candidate weight solution $\boldsymbol{w}_b$. Our developed perturbation function allows the ensemble weight optimization problem to be well-suited for working with the following SA optimization: select a *weight configuration* that will result in the lowest energy (i.e., the best ensemble generalization performance).

We accept the candidate solution $\boldsymbol{w}_b$ with a probability of 1 when the energy $E_{\boldsymbol{w}_b}(\boldsymbol{x}|\boldsymbol{V})$ decreases. On the contrary, when $E_{\boldsymbol{w}_b}(\boldsymbol{x}|\boldsymbol{V})$ increases, we use the following probability:

$$\exp\left[\frac{-\left(E_{\boldsymbol{w}_b}(\boldsymbol{x}|\boldsymbol{V}) - E_{\boldsymbol{w}_{b-1}}(\boldsymbol{x}|\boldsymbol{V})\right)}{T_b}\right] \tag{7}$$

where T_b denotes the annealing temperature at the *b*-th annealing step. If $\exp[\frac{-(E_{\boldsymbol{w}_b}(\boldsymbol{x}|\boldsymbol{V}) - E_{\boldsymbol{w}_{b-1}}(\boldsymbol{x}|\boldsymbol{V}))}{T_b}] > \text{Rand}[0, 1)$, we accept the candidate solution, $\boldsymbol{w}_b$. Here, $\text{Rand}(\,)$ function returns a random integral number between 0 and 1. The aforementioned procedure is repeated many times, which is sufficient to poll all weight values w_k (randomly chosen) several times.

The annealing temperature was initially set to a relatively large value, and then decreased according to the annealing schedule [38]:

$$T_b = 0.9T_{b-1} \tag{8}$$

where *b* is the annealing step and $T_0 = 3E_{\boldsymbol{w}_0}(\boldsymbol{x}|\boldsymbol{V})$. SA runs until the annealing step b reaches the maximum value b_{max} or when the stopping criterion is satisfied (e.g., setting the stopping criterion to the final temperature or energy that must be low enough).

Our proposed SA-based algorithm is summarized in Algorithm 1. To our knowledge, this algorithm has not been previously explored in the field of deep learning with ensemble classification and constitutes the main contribution of our work. Fig. 2 shows that early in the annealing process when the temperature is high, our proposed SA-based algorithm explores a wide range of ensemble weight configurations. Later, as the temperature is lowered, only energy states close to its minimum are tested, and finally, energy is converged to the minimum. It is observed that 140 iterations are generally sufficient for convergence (energy is less than a certain threshold E_{final}). In addition, as shown in Fig. 2(b), *decreasing energy values are strongly correlated with a reduction in the DCNN ensemble generalization error*. This demonstrates that the SA algorithm equipped with our novel energy and perturbation functions is quite useful for finding optimal DCNN ensemble weights (optimal in the sense that the generalization classification error is minimized).

VI. Experiments

An extensive experimental study was carried out to evaluate the effectiveness of the proposed DCNN ensemble construction and SA-based ensemble weight optimization for FER. The following subsections describe the used benchmark datasets that contain images with either real or acted facial expressions in the wild, as well as our evaluation methodology.

Authorized licensed use limited to: HANKUK UNIVERSITY OF FOREIGN STUDIES. Downloaded on January 17,2023 at 18:31:11 UTC from IEEE Xplore. Restrictions apply.

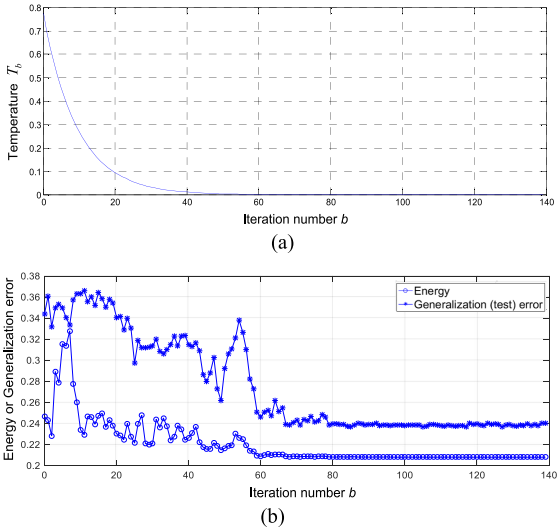


Fig. 2. (a) Annealing schedule of the temperature T_b as a function of iteration number b. (b) The energy $E_{\boldsymbol{w}_b}(\boldsymbol{x}|\boldsymbol{V})$ versus iteration number b, and the associated ensemble generalization (test) error obtained with each candiate weight solution $\boldsymbol{w}_b$. The public test data of the FER2013 DB [29] was used as the validation set to compute both $E_{\boldsymbol{w}_b}(\boldsymbol{x}|\boldsymbol{V})$ and $\boldsymbol{w}_b$, while the private test data was used for computing the ensemble generalization error.

A. Experimental Setup and Condition

The FER2013 [29], SFEW2.0 [53], and RAF-DB [67] datasets collected in the wild were used in our experiments. The FER2013 dataset was created using the Google image search engine to search for images of faces that match a set of 184 emotion-related keywords, such as blissful, enraged, etc. These keywords were combined with words related to gender, age, or ethnicity to obtain nearly 600 strings that were used as facial image search queries. The first 1,000 images returned for each query were kept for the next stage of processing. The images were resized to 48 $\times$ 48 pixels and converted to a grayscale format. The resulting dataset contains 35,887 grayscale images acquired in a wild setting. The SFEW2.0 DB [53] was created by selecting static frames from the acted facial expressions in the wild (AFEW). The SFEW2.0 DB covers unconstrained facial expressions, different head poses, large age ranges, varied focus, occlusions, different resolutions of faces, and close to real-world illuminations. The SFEW2.0 DB is divided into two sets, which are created in a strict person-independent manner. The RAF-DB is a widely used wild FER dataset acquired in an unconstrained setting, offering a broad diversity across pose, gender, age, demography, image quality, and illumination. RAF-DB contains 30,000 facial images annotated with basic or compounded expressions by 40 trained human coders. In our experiments, only images with six discrete basic emotions were used.

Table I shows the number of images for the seven basic expressions in FER2013, SFEW2.0, and RAF-DB. In addition, Figs. 3 and 4 show some image examples for the respective FER DB. As shown in Figs. 3 and 4, face images of our collected datasets are of the great variability in head pose degrees captured in the wild conditions, which is partciulary difficult to

TABLE I
NUMBER OF IMAGES PER EACH EXPRESSION IN USED DATABASES

Expression \ DB	FER2013[29]	SFEW2.0 [53]	RAF-DB [67]
Angry	4,953	104	867
Disgust	547	81	877
Fear	5,121	90	355
Happiness	8,989	112	5,957
Neutral	6,198	98	3,208
Sadness	6,077	92	2,460
Surprise	4,002	86	1,619



Fig. 3. Facial examples of the FER2013 [29]. The columns show the emotion categories.



Fig. 4. Example of facial images (scaled with a size of 224 $\times$ 224 $\times$ 3 pixels) and their original images from (a) SFEW2.0 [53] and (b) RAF-DB [67].

Authorized licensed use limited to: HANKUK UNIVERSITY OF FOREIGN STUDIES. Downloaded on January 17,2023 at 18:31:11 UTC from IEEE Xplore. Restrictions apply.

tackle for correct FER [54], [70]. We measured head pose angles of the face images used in our experiments. For this, we employed the popular "Intraface" toolbox [66] to estimate the three angles, i.e., yaw, pitch, and roll in the degree, of original images (see Fig. 4) of each FER DB. Based on the statistics of our collected data, the range of yaw is $[-43.43, 47.01]$, $[-55.79, 52.48]$, and $[-61.32, 58.72]$ in degrees for FER2013 (test), SFEW2.0 (validation), and RAF-DB (test), respectively. In terms of pitch range, $[-32.32, 45.26]$, $[-15.54, 37.89]$, and $[-43.85, 54.84]$, while $[-20.34, 19.86]$, $[-29.8, 30.22]$, and $[-31.86, 28.41]$ for roll, are observed for the aforementioned order of FER datasets. Moreover, among their original test images, about one-fifth (20%) of FER2013 (test), about half (50%) of SFEW2.0 (validation), and about one-fifth (20%) of RAF-DB (test) have poses larger than 30 degrees (in yaw or pitch or roll). These observations demonstrate that our method is tested under in-the-wild FER scenario, which covers a wide range of pose variations.

As for DCNN ensemble members, we chose VGG-Face [39] as the pre-trained deep network because it performs well and involves a moderate number of parameters. *Note that other recent pre-trained deep CNN or newly designed deep CNN networks [12], [54] can be readily applied to our SA algorithm.* The VGG-Face is a deep CNN model successfully trained on 2.6 million facial images collected from the web to recognize 2,622 identities. This network involves 16 convolutional layers, five max-pooling layers, three fully connected (FC) layers, and a final layer with softmax activation. To construct an ensemble of DCNNs, we used grayscale, LBP, and Gabor face representations as an input to the individual DCNNs. Four different LBP representations were obtained by adjusting the parameter values [33] (no. of sampling points, radius): (8,1), (8,2), (8,3), and (16,2). In addition, referring to [33], 2D Gabor kernels with three different scales (1, 3, and 4 scales) and three orientations (0, 4, and 5 orientations) were used to create nine Gabor representations. Note that we used four bootstrap training datasets [16] and 14 different face representations, resulting in a total of 56 DCNN ensemble members.

We used the training dataset of each FER DB for fine-tuning the VGG-Face model as in [40]–[41] and scaled the size of image data to $224 \times 224 \times 3$ to fit the VGG-Face input requirement. The hyper-parameters of each network are the same as those used in [41]: momentum 0.9; weight decay 5×10^{-4}; and initial learning rate 10^{-2}, which is decreased by a factor of 10 when the validation error stops decreasing (specifically, when the error increases for more than three consecutive times). Overall, each DCNN was trained using three decreasing learning rates. In the proposed SA algorithm, we set $b_{max} = 1,000$, $T_{\text{final}} = 10^{-8}$, $E_{\text{final}} = -\infty$, and $try_{max} = 300$. The λ parameter in (3) was experimental chosen by means of an exhaustive tuning process where λ is varied over the range [0,1], using a step size equal to "0.05". The determination of λ is made by selecting the one having the best FER accuracy on the validation set of each FER dataset. Our results show that a setting λ in the range of [0.25, 0.35] is found to be adequate for all of the datasets used in our experiments. In addition, there is little performance difference in the range of [0.25, 0.35]. For this reason, we set $\lambda = 0.3$ for all the reported results, realizing a good compromise between the accuracy and diversity of the DCNN ensemble.

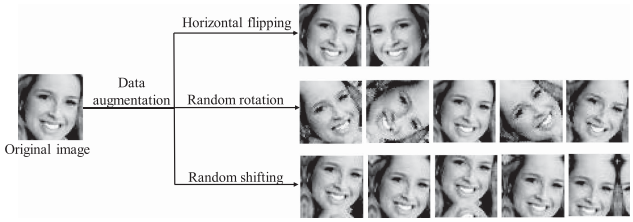


Fig. 5. Examples of randomly augmented facial images with three data augmentation techniques.

Three data augmentation strategies were used during the fine-tuning phase, which included horizontal flips, random rotations (in a degree range of $[-60, 60]$), and random shifting. To implement data augmentation, we used the Keras ImageDataGenerator API.[1] Fig. 5 shows the original facial images and the corresponding randomly augmented sample images.

B. Effectiveness of Proposed SA Algorithm for Finding Optimal DCNN Ensemble Weights

To demonstrate the usefulness of our SA-based algorithm in terms of finding optimal (DCNN) ensemble weights, a comparative experimental study was carried out. For comparison, we employed other popular ensemble weight computation approaches. Specifically, the following four approaches were compared: (a) conventional majority voting [28], (b) performance weighting [26], [31], (c) random search [60], and (d) attention network [65]. In the case of majority voting, all DCNN members have the same combination weights [31] and are treated equally (i.e., this approach is the same as that proposed in conventional majority voting [28]). When using performance weighting [26], [31], we weight each DCNN according to its individual performance on the validation set. The best DCNN can be assigned a weight of one, whereas the weight of the worst DCNN is zero [26], [31]. The DCNN classifiers − whose performances are given weights − are determined as in [26], [31]. For implementing random search [60], uniform DCNN ensemble weights are generated, and then a sampling procedure is performed to find the ensemble weights that yield the highest validation set performance. Furthermore, attention network [65] was used to generate DCNN weights; we first pooled all deep face features, each extracted from the FC layer of an individual DCNN, and then apply them to attention blocks to adaptively compute the weights [65].

From Table II, we observe the following: 1) Ensemble results have been compared with the baseline performance of a best single DCNN classifier. *Every DCNN ensemble combination approach has shown a better performance than a best single DCNN.* 2) DCNN ensemble with different combination weights performs better than the DCNN ensemble combined via naive majority voting [28] with uniform weights, which indicates that *it is beneficial to have an ensemble of DCNNs with different degrees of contributions* to improve performance. 3) the performance improvement is quite convincing when ensemble weights

[1][Online]. Available: https://keras.io/preprocessing/image

Authorized licensed use limited to: HANKUK UNIVERSITY OF FOREIGN STUDIES. Downloaded on January 17,2023 at 18:31:11 UTC from IEEE Xplore. Restrictions apply.

TABLE II
COMPARISON OF FER ACCURACIES WITH RESPECT TO THE DIFFERENT DCNN ENSEMBLE WEIGHT COMPUTATION APPROACHES. THE PRIVATE TEST SET OF THE FER2013 DB WAS USED FOR TESTING. (BOLD: BEST, UNDERLINE: SECOND BEST)

Approach	FER Accuracy (%)
Best individual accuracy (baseline)	65.31
Uniform weighting (majority voting) [28]	67.56
Performance weighting [26, 31]	71.15
Weighting using attention network [65]	73.97
Weighting using random search [60]	<u>74.68</u>
Weighting using proposed SA-based algorithm	**76.69**

∗SA: Simulated annealing

TABLE III
COMPARISON OF THE DIVERSITIE MEASURES [13] AMONG THE DCNN MEMBERS FOR THREE DIFFERENT ENSEMBLE CONSTRUCTION APPROACHES

Diversity measure	DCNN ensemble construction approach		
	Using only grayscale face representation with bagging	Using different face representations without bagging	Using different face representations with bagging
Q statistic	0.862	0.421	0.042
Correlation coefficient	0.692	0.354	-0.193
Double-default measure	0.835	0.419	0.218

determined by our SA algorithm are used. In particular, our SA algorithm is superior to the approach of choosing ensemble weights based on the attention network proposed in [65]. The underlying reason for this is that we find and adopt ensemble weights by **directly minimizing the estimate of the DCNN ensemble generalization error via our SA algorithm**.

C. Demonstrating Our DCNN Ensemble Construction for Increasing Diversity

To assess the effectiveness of our ensemble construction method in improving ensemble diversity, we used the following three pairwise diversity measures [13]: Q statistic, correlation coefficients, and double-default measure. The possible ranges of Q statistic, correlation coefficients, and double-default measure are $[-1\ 1]$, $[-1\ 1]$, and $[0\ 1]$, respectively, for a pair of individual DCNN members. Note that the lower the values of the aforementioned diversity measures, the more diverse the ensemble members are.

Table III shows the three diversity measures among the DCNN members with respect to three different ensemble construction approaches. Note that each diversity measure was averaged over a total number of different DCNN member pairs. As shown in Table III, for all types of diversity measures, the DCNN ensemble trained by our approach (using different face representations with bagging) was found to be better than its counterparts for producing more diverse ensemble members. Especially, our approach achieved a negative value of the correlation coefficient; this indicates that DCNN members tend to be negatively correlated. In general, combining classifiers are most effective when the errors in the individual classifiers are negatively correlated [44]. Overall, the results validate the advantage of using different face representations with bagging to produce a more diverse (uncorrelated) DCNN ensemble.

TABLE IV
FER ACCURACY COMPARISONS WITH OTHER STATE-OF-THE-ART APPROACHES ON FER2013 PRIVATE TEST DATASET. RESULTS FOR THE COMPARISON ARE DIRECTLY CITED FROM PAPERS RECENTLY PUBLISHED. (BOLD: BEST, UNDERLINE: SECOND BEST)

Approach		FER accuracy(%)
Tang 13 [46]	DCNN using L2-SVM loss (baseline)	71.16
Kim *et al.* 16 [5]	DCNN ensemble in a hierarchical committee	72.72
Pramerdorfer *et al.* 16 [59]	DCNN ensemble with multiple deep architectures	75.20
Connie *et al.* 17 [47]	Hybrid CNN–SIFT aggregator	73.40
Li *et al.* 18 [42]	MRMR-based ensemble prunning	70.66
Shao *et al.* 19 [19]	Pre-trained CNN with ResNet101	71.14
Gan *et al.* 19 [48]	DCNN ensemble with soft label construction	73.73
Georgescu *et al.* 19 [45]	Local learning with a combination of deep and handcrafted features	<u>75.42</u>
Proposed method	DCNN ensemble with multiple input representations and SA-based ensemble weight optimization	**76.69**

D. Comparison With State-of-the-art FER Methods

In this section, we compare the proposed method for in-the-wild facial expression against other state-of-the-art methods. Note that a comparative study done by implementing other methods for ourselves may not guarantee fair and stable comparison, as we are likely to miss some optimal parameters and incorrectly perform their tuning process. To circumvent this problem, we make direct comparisons with other state-of-the-art results recently reported by other researchers on the FER2013 [29], SFEW2.0 [53], and RAF-DB [67]. Therefore, all the results of the comparison are directly cited from recently published papers.

The FER2013 consists of 28,709 training, 3,589 public test, and 3,589 private test images under seven different types of emotions. Note that a training set was used to create DCNN ensemble members, while 3,589 public test images were used as the validation set for optimizing their ensemble weights, and 3,589 private test images were used as the testing set for performance evaluation. Table IV lists the comparative results of the FER2013 DB. Compared to other state-of-the-art approaches, our proposed method achieves the highest recognition performance of 76.69%, outperforming the most recent best performance [45], i.e., 75.42%, on the FER2013 DB. In addition, comparative experiments were performed on SFEW2.0, released for a sub-competition in the 3rd Emotion Recognition in the Wild 2015 (EmotiW2015) challenge [53]. The SFEW2.0 version targets the efforts required toward an effect analysis in wild

Authorized licensed use limited to: HANKUK UNIVERSITY OF FOREIGN STUDIES. Downloaded on January 17,2023 at 18:31:11 UTC from IEEE Xplore. Restrictions apply.

TABLE V
FER ACCURACY COMPARISONS ON SFEW2.0. RESULTS FOR THE COMPARISON ARE DIRECTLY CITED FROM PAPERS RECENTLY PUBLISHED. VALIDATION ACCURACY WAS REPORTED

Approach		FER accuracy(%)
Dhall *et al.* 11 [53]	SVM based on the fusion of handcrafted features (baseline)	35.93
Yu *et al.* 15 [4]	Deep network ensemble using different loss-based optimization frameworks	56.19
Kim *et al.* 15 [62]	Deep network ensemble using a hierarchical committee	53.90
Liu *et al.* 17 [56]	(*N*+*M*)-tuples clusters loss	54.19
Ding *et al.* 17 [57]	FaceNet2ExpNet based on two-stage training	55.15
Meng *et al.* 17 [58]	Identity-aware convolutional neural network	50.98
Acharya *et al.* 18 [63]	Manifold network with covariance pooling	58.20
Zeng *et al.* 18 [64]	LTNet with inconsistent pseudo annotations to latent truth (IPA2LT)	58.29
Ly *et al.* 19 [61]	Multimodal 2D and 3D facial features	56.90
Gan *et al.* 19 [48]	DCNN ensemble with soft label construction	55.73
Proposed method	DCNN ensemble with multiple input representations and SA-based ensemble weight optimization	**58.68**

FER. The SFEW2.0 consists of three sets: train (958 samples), val (436 samples), and test (372 samples). Each facial image is assigned to one of the seven expression categories (angry, disgust, fear, sadness, happiness, surprise, and neutral). The expression labels of the training and validation sets are publicly available, whereas those of the testing set are held back by the challenge organizer. To cope with a small number of training samples (only 958 images) when creating DCNN members in our method, data augmentation techniques were applied for the SFEW2.0 training set. We generated 11 augmentations per facial image, as described in Fig. 5, yielding 11,496 images as the augmented training dataset. Note that because the labels of the test dataset were not provided, the validation set was used as the test set for the experiment in this work, while 4,598 facial images (around 40%) from our augmented training dataset were randomly selected to become a part of our validation dataset. Table V shows the comparison between the FER performances of our proposal with those of other approaches. Because we do not have access to the test data, we report the results on the validation data. The proposed method outperforms the baseline [53] of SFEW2.0 by a large margin of 22.67%. Our method attains a 58.68% FER accuracy, achieving competitive performance with the state-of-the-art results of 58.20% [63] and 58.29% [64].

To further validate whether the proposed approach generalized well with another wild FER dataset acquired in an unconstrained setting, images with six discrete basic emotions were collected from the RAF-DB [67], including 12,271 images as training data and 3,068 images as test data. For constructing a validation dataset, 4,800 images from the full training (i.e., 12,271 images) set were randomly chosen. As shown in Table VI, the FER performance of our method on RAF-DB, i.e., 87.13%, is quite comparable with the most recent best performance, i.e., 87.78%, reported in [71].

TABLE VI
FER ACCURACY COMPARISONS ON RAF-DB. RESULTS FOR THE COMPARISON ARE DIRECTLY CITED FROM PAPERS RECENTLY PUBLISHED

Approach		FER accuracy(%)
Li *et al.* 19 [68]	Convolutional neural network with attention mechanism	85.07
Wang *et al.* 20 [69]	Self-Cure Network (SCN)	87.03
Wang *et al.* 20 [70]	Region Attention Network (RAN)	86.90
Momin *et al*. 21 [72]	Deep network ensemble using multi-architectural representations	82.07
Wang *et al*. 21 [73]	Oriented Attention Ensemble Network (OAENet)	86.50
Farzaneh *et al.* 21 [71]	Deep Attentive Center Loss (DACL)	**87.78**
Proposed method	DCNN ensemble with multiple input representations and SA-based ensemble weight optimization	87.13

TABLE VII
CONFUSION MATRIX (UNIT OF %) OF THE PROPOSED DCNN ENSEMBLE METHOD EVALUATED ON THE SFEW 2.0 VALIDATION SET

	Ang	Dis	Fea	Hap	Neu	Sad	Sur
Ang	**67.30**	0	3.08	2.03	4.49	12.86	10.24
Dis	20.09	**11.36**	47.27	5.81	0	3.63	11.84
Fea	19.03	40.41	**14.44**	2.08	3.78	18.37	1.89
Hap	2.50	0	0.91	**84.52**	8.98	2.18	0.91
Neu	7.50	0	6.86	2.39	**60.54**	22.56	0.15
Sad	15.11	1.65	10.48	2.53	9.97	**50.06**	10.20
Sur	1.20	2.11	11.36	6.31	2.16	0.48	**76.38**

Results in Tables IV, V, and VI show that FER accuracy of our method is quite comparable or better than the most recent best performance reported in [45], [64], [71]. This demonstrates that our proposed DCNN ensemble approach has a *better discrimination ability for expression characteristics*, as well as a *better generalization ability* in terms of attaining competitive FER performances compared with the state-of-the-art methods on various and different wild FER datasets.

VII. DISCUSSION AND CONCLUSION

The confusion matrix of our proposed DCNN ensemble is reported in Table VII. We can see that the accuracy for "disgust" and "fear" are much lower than other expressions, which is consistent with results previously reported for the SFEW2.0 datasets [4], [61]–[64]. This may be attributed to the fact that such emotion class is inherently more challenging to classify or simply because of the relatively small number of samples available for each emotion class (see Table I). However, note that the proposed method considerably improves the accuracy on "disgust" or "fear" when compared to the corresponding results reported by other works [4], [61]–[64].

Authorized licensed use limited to: HANKUK UNIVERSITY OF FOREIGN STUDIES. Downloaded on January 17,2023 at 18:31:11 UTC from IEEE Xplore. Restrictions apply.

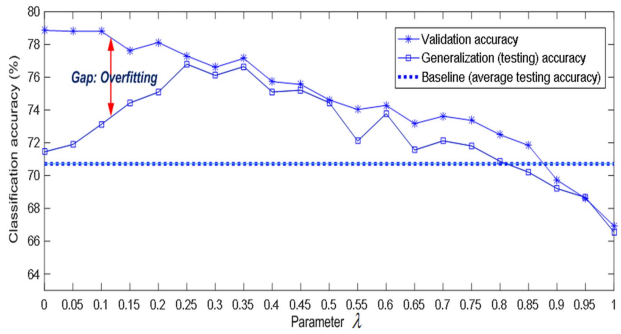


Fig. 6. Impact of parameter λ on generalization (testing) performance in our SA algorithm. Validation and testing accuracies were computed using the FER2013 public and private data sets, respectively. Note that the gap between validation and generalization performance represents the degree of overfitting.

The runtime of our DCNN ensemble classifier is approximately 85 ms per image on a hardware configuration comprising of Intel Xeon E5-2620 v4 CPUs, 256-GB RAM, when accounting for facial image transformation (into LBP and/or Gabor face representations), the feedforward network, and the final combined classification output. *This level of runtime performance should be reasonable for many practical FER applications as almost 12 facial images can be processed per second for FER*. In summary, the proposed method can achieve better recognition accuracy than other state-of-the-art results on three wild FER DBs, in practically efficient runtime performance. This indicates that the proposed DCNN ensemble approach can be used as a competitive solution for expression recognition in the wild.

Note that the parameter λ shown in (3) was devised to incorporate the diversity term into the energy function to be minimized. We demonstrate that *properly adjusting λ can be beneficial for improving the generalization (testing) performance*. The results in Fig. 6 justify the advantage of using λ (i.e., the ensemble diversity term). As shown in Fig. 6, the validation accuracy (obtained using the validation set used to optimize DCNN ensemble weights) reaches a maximum for $\lambda = 0$, while the generalization accuracy is lower and smaller than the averaged generalization accuracy along with λ. Note that $\lambda = 0$ means that the ensemble accuracy term $\mathrm{Acc}_{\boldsymbol{w}}(\cdot)$ in (3) is only considered as an energy function in our SA algorithm. **For generalization accuracy, the maximum is achieved in the range of** $[0.25, 0.35]$ **for λ**. The rationale behind this observation is that for $\lambda = 0$, ensemble weights are adjusted in such way that the DCNN ensemble classifier is strongly overfitted to the validation set so that the diversity would be quite small. In contrast, as λ approximates to 1, diversity would be facilitated at the cost of ensemble accuracy. This observation is consistent with the commonly held idea that the benefit of an ensemble classifier would be maximized when an optimal balance between classification accuracy and diversity is found [13], [32]. Consequently, the use of λ for the proper adjustment of ensemble accuracy and diversity is beneficial to achieve a low generalization error.

Prior to concluding this paper, we presented feature activation map [75] of a test face image for examining the internal representations learned by our DCNN ensemble. For this purpose, feature maps from the VGG-Face networks, each trained using one of the nine Gabor face representations with three scales (1, 3, and 4 scales) and three orientations (0, 4, and 5 orientations), were obtained from the conv2_1 layer of VGG-Face network [39]; note that the strongest activations for this selected layer were displayed. As can be observed in Fig. 7, our DCNN ensemble captures **complementary (diverse) activation patterns from various input (Gabor) face representations**, and therefore they are likely to be mutually compensational in terms of extracting different deep face features of an input face image. This allows DCNN ensemble members to be diverse (to a large extent), leading to **highly complementary deep face features** for correct FER. The observations in Fig. 7 is line with our argument that applying different face representations for the construction of DCNN ensemble is useful to induce diversity among DCNN members.

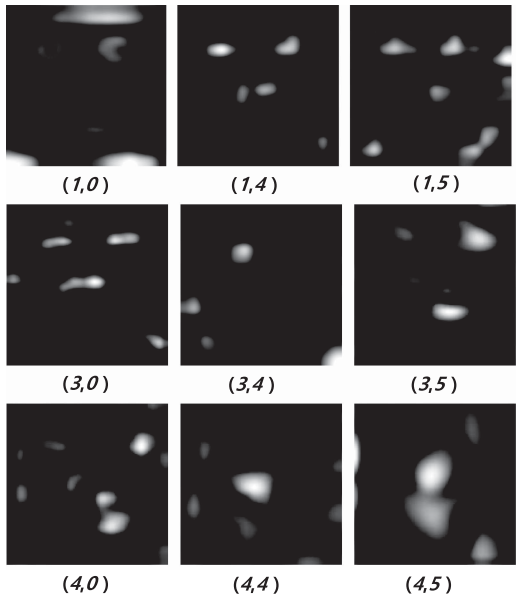


Fig. 7. Visualization of feature activation maps [75] of a test face image from the VGG-Face networks [39] as ensemble members, each trained using one of the nine different Gabor face representations with three scales and three orientations. The strongest activations for the selected conv2_1 layer are displayed. An associated (*scale*, *orientation*) of the Gabor kernel is described below each corresponding feature map.

For improving FER in the wild, we proposed the DCNN ensemble classification approach. In particular, we developed a novel simulated annealing (SA) based algorithm that exploits optimal ensemble weights by minimizing the energy function composed of ensemble accuracy and diversity terms. We have shown that our SA-based algorithm can be used to find optimal DCNN ensemble weights, which can lead to a considerably better generalized recognition performance than other popular ensemble weight computation methods. In addition, our proposed way of combining different face representations and bagging for the creation of DCNN ensemble is shown to be effective in terms of increasing diversity among the outputs of individual DCNNs. The results show that our DCNN ensemble approach can achieve competitive performances on challenging wild FER datasets in both recognition accuracy and runtime.

Authorized licensed use limited to: HANKUK UNIVERSITY OF FOREIGN STUDIES. Downloaded on January 17,2023 at 18:31:11 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] D. CireşAn, U. Meier, J. Masci, and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification," *Neural Netw.*, vol. 32, pp. 333–338, 2012.

[2] X. S. Wei, B. B. Gao, and J. Wu, "Deep spatial pyramid ensemble for cultural event recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2015, pp. 38–44.

[3] S. E. Kahou *et al.*, "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[4] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2015, pp. 435–442.

[5] B. K. Kim, J. Roh, S. Y. Dong, and S. Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, 2016.

[6] R. C. Malli, M. Aygun, and H. K. Ekenel, "Apparent age estimation using ensemble of deep learning models," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 714–721.

[7] H. Siqueira, S. Magg, and S. Wermter, "Efficient facial feature learning with wide ensemble-based convolutional neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 5800–5809.

[8] Y. D. Kim, T. Jang, B. Han, and S. Choi, "Learning to select pre-trained deep representations with bayesian evidence framework published in computer vison and pattern recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5318–5326.

[9] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, vol. 4, no. 1, pp. 1–58, 1992.

[10] G. Brown, J. L. Wyatt, and P. Tino, "Managing diversity in regression ensembles," *J. Mach. Learn. Res.*, vol. 6, pp. 1621–1650, 2005.

[11] G. Brown, *Diversity in Neural Network Ensembles*, Ph.D. dissertation, School Comput. Sci., Birmingham Univ., Birmingham, U.K., 2004.

[12] H. Li, J. Sun, Z. Su, and L. Chen, "Multimodal 2D+3D facial expression recognition with deep fusion convolutional neural network," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2816–2831, Dec. 2017.

[13] L. I. Kuncheva and C. J. Whitaker, "Measure of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Mach. Learn.*, vol. 51, no. 2, pp. 181–207, 2003.

[14] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. LNCS Int. Workshop Mult. Classifier Syst.*, 2000, pp. 1–15.

[15] A. Mollahosseini *et al.*, "Facial expression recognition from world wild web," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 58–65.

[16] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Mach. Learn.*, vol. 40, no. 2, pp. 139–157, 2000.

[17] R. Ranawana and V. Palade, "Multi-classifier systems—A review and roadmap for developers," *Inf. Sci.*, vol. 179, no. 2, pp. 1298–1318, 2009.

[18] J. Joo, W. Li, F. F. Steen, and S. C. Zhu, "Visual persuasion: Inferring communicative intents of images," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 216–223.

[19] J. Shao and Y. Qian, "Three convolutional neural network models for facial expression recognition in the wild," *Neurocomputing*, vol. 355, pp. 82–92, 2019.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[21] T. Gehrig and H. K. Ekenel, "Why is facial expression analysis in the wild challenging?," in *Proc. ACM Emotion Recognit. Wild Challenge Workshop*, 2013, pp. 9–16.

[22] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image Vis. Comput.*, vol. 47, pp. 3–18, 2016.

[23] Y. Ren, L. Zhang, and P. N. Suganthan, "Ensemble classification and regression - Recent developments, applications and future directions," *IEEE Comput. Intell. Mag.*, vol. 11, no. 1, pp. 41–53, Feb. 2016.

[24] D. W. Optiz and J. W. Shavlik, "Generating accurate and diverse members of a neural-network ensemble," in *Proc. Adv. Neural Inf. Process. Syst.*, 1996, pp. 531–541.

[25] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 1995, pp. 231–238.

[26] L. Rokach, "Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography," *Comput. Statist. Data Anal.*, vol. 53, no. 12, pp. 4046–4072, 2009.

[27] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, 2015.

[28] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[29] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.*, 2013, pp. 117–124.

[30] M. P. Perrone and L. N. Cooper, "When networks disagree: Ensemble methods for hybrid neural networks," in *Neural Networks for Speech and Image Processing*, R. J. Mammone, Ed. New York, NY, USA: Chapman & Hall, 1993, pp. 126–142.

[31] Z. H. Zhou, *Ensemble Methods: Foundation and Algorithms*, Boca Raton, FL, USA: CRC, 2012.

[32] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: A survey and categorization," *Inf. Fusion*, vol. 6, no. 1, pp. 5–20, 2005.

[33] J. Y. Choi, Y. M. Ro, and K. N Plataniotis, "Color local texture features for color face recognition," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1366–1380, Mar. 2012.

[34] R. McKay and H. Abbass, "Anticorrelation measures in genetic programming," in *Proc. Australasia-Jpn. Workshop Intell. Evol. Syst.*, 2001, pp. 45–51.

[35] N. Ueda and R. Nakano, "Generalization error of ensemble estimators," in *Proc. IEEE Int. Conf. Neural Netw.*, 1996, pp. 90–95.

[36] R. S. Sexton, R. E. Dorsey, and J. D. Johnson, "Beyond backpropagation: Using simulated annealing for training neural networks," *J. Organizational End User Comput.*, vol. 11, no. 3, pp. 3–10, 1999.

[37] R. A. Rutenbar, "Simulated annealing algorithms: An overview," *IEEE Circuits Devices Mag.*, vol. 5, no. 1, pp. 19–26, Jan. 1989.

[38] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed.. Hoboken, NJ, USA: Wiley, 1973.

[39] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *Brit. Mach. Vis.*, vol. 1, no. 3, 2015, Art. no. 6.

[40] A. Vedaldi and K. Lenc. "MatConvNet—Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 689–692.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.

[42] D. Li and G. Wen, "MRMR-based ensemble pruning for facial expression recognition," *Multimed Tools Appl.*, vol. 77, pp. 15251–15272, 2018.

[43] A. Pentina and R. Urner, "Lifelong learning with weighted majority votes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3612–3620.

[44] H. Chen and X. Yao, "Multiobjective neural network ensembles based on regularized negative correlation learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 12, pp. 1738–1751, Dec. 2010.

[45] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, May 2019.

[46] Y. Tang, "Deep learning using linear support vector machines," 2013, *arXiv:1306.0239*.

[47] T. Connie, M. Al-Shabi, W. P. Cheah, and M. Goh, "Facial expression recognition using a hybrid CNN–SIFT aggregator," in *Proc. Int. Workshop Multi-Disciplinary Trends Artif. Intell.*, 2017, pp. 139–149.

[48] Y. Gan, J. Chen, and L. Xu, "Facial expression recognition boosted by soft label with a diverse ensemble," *Pattern Recognit. Lett.*, vol. 125, pp. 105–112, 2019.

[49] G. Ali, M. A. Iqbal, and T. S. Choi, "Boosted NNE collections for multicultural facial expression recognition," *Pattern Recognit.*, vol. 55, pp. 14–27, 2016.

[50] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, 2017.

[51] S. Zafeiriou, A. Papaioannou, I. Kotsia, M. Nicolaou, and G. Zhao, "Facial affect 'In-The-Wild': A survey and new database," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 36–47.

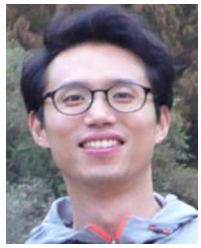
[52] M. Ghayoumi, "A quick review of deep learning in facial expression," *J. Commun. Comput.*, vol. 14, pp. 34–38, 2017.

[53] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 2106–2112.

[54] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2020.2981446.

Authorized licensed use limited to: HANKUK UNIVERSITY OF FOREIGN STUDIES. Downloaded on January 17,2023 at 18:31:11 UTC from IEEE Xplore. Restrictions apply.

[55] M. Liu, S. Li, S. Shan, and X. Chen, "Au-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126–136, 2015.
[56] X. Liu, B. Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 20–29.
[57] H. Ding, S. K. Zhou, and R. Chellappa, "FaceNet2expNet: Regularizing a deep face recognition net for expression recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 118–126.
[58] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 558–565.
[59] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: State of the art," 2016, *arXiv:1612.02903*.
[60] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012.
[61] T. S. Ly, N.T. Do, S.H. Kim, H. J. Yang, and G. S. Lee, "A novel 2D and 3D multimodal approach for in-the-wild facial expression recognition," *Image Vis. Comput.*, vol. 92, no. 103817, pp. 1–12, 2019.
[62] B. K. Kim, H. Lee, J. Roh, and S. Y. Lee, "Hierarchical committee of deep CNNs with exponentially-weighted decision fusion for static facial expression recognition," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2015, pp. 427–434.
[63] D. Acharya, Z. Huang, D. Paudel, and L. V. Gool, "Covariance pooling for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 480–487.
[64] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 222–237.
[65] J. Yang *et al.*, "Neural aggregation network for video face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4362–4371.
[66] F. D. L. Torre *et al.*, "*IntraFace*," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2015, pp. 1–8.
[67] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,"in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2584–2593.
[68] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.
[69] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6896–6905.
[70] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.
[71] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2021, pp. 2402–2411.
[72] R. Momin, A. S. Momin, and K. Rasheed, "Recognizing facial expressions in the wild using multi-architectural representations based ensemble learning with distillation," 2021, *arXiv:2106.16126*.
[73] Z. Wang, F. Zeng, S. Liu, and B. Zeng, "OAENet: Oriented attention ensemble for accurate facial expression recognition," *Pattern Recognit.*, vol. 112, 2021, Art. no. 107694.
[74] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.



Jae Young Choi (Member, IEEE) received the M.S. and Ph.D. degrees from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2008 and 2011, respectively. In 2008, he was a Visiting Scholar with the University of Toronto, Toronto, ON, Canada, and from 2011 to 2012, he was a Postdoctoral Researcher with the University of Toronto. From 2012 to 2013, he was a Postdoc Fellow with the University of Pennsylvania, Philadelphia, PA, USA. From 2013 to 2014, he was a Senior Engineer with Samsung Electronics. He is currently an Associated Professor with the Division of Computer Engineering, Hankuk University of Foreign Studies, Seoul, South Korea. He is the author or coauthor of more than 100 refereed research publications in his research field, which include deep learning, ensemble machine learning, pattern recognition, and computer vision. Especially, he has developed several pioneering algorithms for automatic face recognition using facial color information. Prof. Choi was the recipient of the Best Paper Award of Korea Multimedia Society in 2021. He was also the recipient of the Samsung HumanTech Thesis Prize in 2010.



Bumshik Lee (Member, IEEE) received the B.S. degree in electrical engineering from Korea University, Seoul, South Korea, and the M.S. and Ph.D. degrees in information and communications engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. He was a Research Professor with KAIST, in 2014 and a Postdoctoral Scholar with the University of California San Diego, San Diego, CA, USA, from 2012 to 2013. From 2015 to 2016, he was a Principal Engineer with Advanced Standard R&D Lab., LG Electronics, Seoul, South Korea. He is currently an Associated Professor with the Department of Information and Communication Engineering, Chosun University, Gwangju, South Korea. His research interests include pattern recognition, video compression and processing, video security, and medical image processing.

Authorized licensed use limited to: HANKUK UNIVERSITY OF FOREIGN STUDIES. Downloaded on January 17,2023 at 18:31:11 UTC from IEEE Xplore. Restrictions apply.