Confidence-Calibrated Face and Kinship Verification

Min Xu, Ximiao Zhang, and Xiuzhuang Zhou

Abstract—In this paper, we investigate the problem of prediction confidence in face and kinship verification. Most existing face and kinship verification methods focus on accuracy performance while ignoring confidence estimation for their prediction results. However, confidence estimation is essential for modeling reliability and trustworthiness in such high-risk tasks. To address this, we introduce an effective confidence measure that allows verification models to convert a similarity score into a confidence score for any given face pair. We further propose a confidence-calibrated approach, termed Angular Scaling Calibration (ASC). ASC is easy to implement and can be readily applied to existing verification models without model modifications, yielding accuracy-preserving and confidence-calibrated probabilistic verification models. In addition, we introduce the uncertainty in the calibrated confidence to boost the reliability and trustworthiness of the verification models in the presence of noisy data. To the best of our knowledge, our work presents the first comprehensive confidence-calibrated solution for modern face and kinship verification tasks. We conduct extensive experiments on four widely used face and kinship verification datasets, and the results demonstrate the effectiveness of our proposed approach. Code and models are available at https://github.com/cnulab/ASC.

Index Terms—Kinship verification, face verification, confidence estimation, confidence calibration, uncertainty estimation.

I. INTRODUCTION

FACE and kinship verification are two popular facial analysis tasks in computer vision. Face verification attempts to determine whether a given pair of face samples belongs to the same individual [1]. Kinship verification aims to predict the existence of a kin relation between a given pair of face samples [2]. Over the past decade, face and kinship verification tasks have achieved remarkable advancements attributed to the application of deep learning technologies, particularly under controlled conditions. However, the two tasks in the wild are still challenging due to large intra-class variations, such as complex background, occlusion, and a variety of variations on illumination, pose and facial expression. Prior works on face and kinship verification are generally devoted to improve prediction accuracy, while paying little attention to confidence estimation. We contend that the reliability is also a key measure for evaluating the performance of these verification algorithms, and it becomes even more crucial for those verification systems deployed in high-risk scenarios. In this paper, we focus on modeling confidence estimation and calibration for face and kinship verification.

Accurate confidence estimation for face and kinship verification often presents a challenge. One reason for this can be the absence of labels that describe sample (pairs) uncertainty in most face or kinship verification datasets. To address this issue, existing methods employ quality networks [3], [4] or uncertainty estimation [5], [6] to assist the decision-making. However, these approaches are more concerned with the quality of individual face images rather than the verification tasks. Additionally, modern neural networks tend to be over- or under-confident in their predictions. Hence, the estimated confidence of a face pair might not accurately reflect the verification accuracy. To overcome these challenges, we first propose a simple yet effective confidence measure for face and kinship verification. Our approach is inspired by the observation that if the pair similarity is close to the decision threshold, the model is less likely to make a correct prediction. Conversely, if the pair similarity is far from the threshold, the model’s prediction is more likely to be correct. The proposed measure allows face and kinship verification models to convert the similarity score into a confidence score for a given face pair. We further develop a new algorithm to adjust the similarity of face pairs in angular space for confidence calibration, so that the calibrated confidence can well quantify the prediction confidence while maintaining the verification accuracy. Furthermore, we integrate the proposed confidence measure and confidence calibration methods with uncertainty estimation, which allows the confidence score to take into account both image quality and the verification process, providing a more reliable reference for decisions.

The pipeline of our confidence-calibrated face and kinship verification is illustrated in Fig. 1. The main contributions of this work can be outlined below:

- We introduce a new confidence measure for probabilistic face and kinship verification, allowing any off-the-shelf verification models to efficiently estimate the prediction confidence.
- We develop a confidence-calibrated algorithm via angle scaling, which can be directly applied to existing face and kinship verification models without model modifications. To the best of our knowledge, this is the first general confidence-calibrated solution to face and kinship verification in a modern context.
The introduction of Large-Margin Softmax [11] sparked new ideas for face verification, and led to several notable studies, such as SpereFace [12], CosFace [13], and ArcFace [14]. These methods penalized the angles between deep features and their corresponding weights in angular space, effectively improving the discriminative ability of feature embeddings. Note that the angular-based distance measure is gradually replacing the Euclidean distance measure as the most popular method for face verification tasks, since the cosine of the angle and the softmax loss are inherently consistent.

The facial kinship verification problem dates back to the pioneering work by Fang et al. [2]. Since then, a variety of approaches have been proposed [15], [16]. These approaches can be roughly divided into two categories: traditional methods based on feature or metric learning, and deep learning-based approaches developed in recent years. The former category seeks to learn a feature encoder or distance metric from pairwise kinship samples. A representative work of this line was the neighborhood repulsed metric learning (NRML) [17], which aimed to learn a distance metric that maximized the distance between negative pairs while minimizing the distance between positive pairs. Following this, [18], [19], [20], [21], [22], [23] expanded on this concept by developing the novel approaches combining multiple features [18], [21], multiple views [20], [21], [22], multiple similarity [19], [22], and denoising [23]. More recently, the traditional methods [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35] have been innovatively enhanced by incorporating deep learning techniques, such as deep metric learning, deep feature representation, or their combination to tackle the kinship recognition problem [24], [27], [30], [31]. Moreover, advanced kinship recognition solutions have emerged that leverage generative modeling [36], [37], [38] or graph representation [39] to further bolster the robustness of kinship verification.

As discussed above, most existing face and kinship verification methods focus on accuracy performance, while overlooking the importance of confidence estimation in their predictions. Even though a few attempts have been made recently to estimate the prediction confidence [40], the estimated confidence can often be inaccurate due to the inadequate calibration of modern DNNs [41]. Differently, our method provides well-calibrated prediction confidence while maintaining the verification accuracy.

II. RELATED WORKS

A. Face and Kinship Verification

Face verification is a well-studied research topic in computer vision. In early studies, the low-level handcrafted features were used for verification and identification. The advent of deep neural networks in the past decade greatly enhances the performance of face verification tasks. Pioneering works such as DeepFace [7] and DeepID [8] introduced deep CNNs into face recognition. These works used deep models to learn effective high-level features, which significantly boosted face recognition performance. FaceNet [9] proposed the use of the triplet loss to learn a direct mapping from face images to a compact Euclidean space, where distances correspond directly to a measure of face similarity. Wen et al. [10] proposed a center loss to enhance the compactness of intra-class samples by minimizing the Euclidean distance between deep features and their corresponding class centers, while combining a softmax loss to guarantee inter-class differences. The introduction of Large-Margin Softmax [11] sparked new ideas for face verification, and led to several notable studies, such as SpereFace [12], CosFace [13], and ArcFace [14]. These methods penalized the angles between deep features and their corresponding weights in angular space, effectively improving the discriminative ability of feature embeddings. Note that the angular-based distance measure is gradually replacing the Euclidean distance measure as the most popular method for face verification tasks, since the cosine of the angle and the softmax loss are inherently consistent.

The facial kinship verification problem dates back to the pioneering work by Fang et al. [2]. Since then, a variety of approaches have been proposed [15], [16]. These approaches can be roughly divided into two categories: traditional methods based on feature or metric learning, and deep learning-based approaches developed in recent years. The former category seeks to learn a feature encoder or distance metric from pairwise kinship samples. A representative work of this line was the neighborhood repulsed metric learning (NRML) [17], which aimed to learn a distance metric that maximized the distance between negative pairs while minimizing the distance between positive pairs. Following this, [18], [19], [20], [21], [22], [23] expanded on this concept by developing the novel approaches combining multiple features [18], [21], multiple views [20], [21], [22], multiple similarity [19], [22], and denoising [23]. More recently, the traditional methods [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35] have been innovatively enhanced by incorporating deep learning techniques, such as deep metric learning, deep feature representation, or their combination to tackle the kinship recognition problem [24], [27], [30], [31]. Moreover, advanced kinship recognition solutions have emerged that leverage generative modeling [36], [37], [38] or graph representation [39] to further bolster the robustness of kinship verification.

As discussed above, most existing face and kinship verification methods focus on accuracy performance, while overlooking the importance of confidence estimation in their predictions. Even though a few attempts have been made recently to estimate the prediction confidence [40], the estimated confidence can often be inaccurate due to the inadequate calibration of modern DNNs [41]. Differently, our method provides well-calibrated prediction confidence while maintaining the verification accuracy.
PFE [5] marked the initiation of integrating data uncertainty estimation into face recognition, achieving this by treating face embedding as a probabilistic embedding and determining the likelihood of a face pair belonging to the same identity. SER-FIQ [6] capitalized on the Bayesian approximation of model uncertainty using dropout as a mechanism to gauge image quality. Building on this, Xie et al. [4] and Liu and Tan [3] put forth a method that involved training either an independent network or an auxiliary network branch to derive the quality scores of the images. The study represented in [48] introduced a learning scheme for the relative uncertainty of a face pair via an extra network branch, with the obtained uncertainty acting as a weight to combine features from two distinct labels. Moreover, there have been concerted efforts to embed uncertainty analysis into face representation learning [49], [50].

Existing methods for modeling uncertainty in facial feature learning only provide an uncalibrated uncertainty estimation during the prediction stage, failing to accurately represent the confidence of the prediction. For verification tasks, our proposed confidence measure and calibration approach can provide a well-calibrated prediction confidence in the decision-making process, providing a direct representation of the prediction’s probability of accuracy, thus aiding in risk assessment. The most closely related work to ours is the approach proposed in [40], where uncertainty of the similarity score is estimated and propagated to the prediction confidence in face verification. However, this approach differs from our approach in the definition of the confidence measure, and it doesn’t take into account the confidence calibration, leading to inaccurate confidence estimation. Experiment results on four widely-used datasets demonstrate the effectiveness of our proposed post-calibration method in face and kinship verification, particularly with respect to the uncertainty metric.

III. OUR APPROACH
A. Preliminaries: Face and Kinship Verification

Given a face pair \((X, Y)\), we define the face embeddings as \(x = f(X), y = f(Y)\), where \(f(\cdot)\) is the feature embedding function often represented by modern DNNs. Existing face and kinship verification methods typically compute a cosine similarity, \(s(x, y)\), as a similarity metric:

\[
  s(x, y) = \frac{x \cdot y}{||x|| ||y||} \quad (1)
\]

The decision function \(g(s, \tau)\) for face and kinship verification can be defined by:

\[
g(s, \tau) = \begin{cases} 
1, & s \geq \tau \\
-1, & s < \tau 
\end{cases} \quad (2)
\]

where \(\tau\) is a predefined threshold that is often empirically set based on the Receiver Operating Characteristic (ROC) curve on a held-out validation set.

Recent works on face or kinship verification show that verification systems built even on popular DNNs may not work reliably, especially when the face images are partially occluded or of low resolution. Therefore, confidence estimation for face and kinship verification plays a key role in such a safety-critical task. However, most existing verification methods based on similarity measure fail to quantify the prediction confidence of face pairs, as the similarity score itself does not exactly reflect the prediction confidence. To address this issue, we propose a simple and flexible confidence measure to quantify the prediction confidence for any face and kinship verification models. Specifically, we estimate the prediction confidence based on the similarity \(s\) and threshold \(\tau\), and then calibrate the confidence in angular space so that the calibrated confidence is directly related to the probability of the prediction being correct.

B. Confidence Measure

Intuitively, if the similarity score of a face pair is equal to the left or right boundary values (−1 or 1), the prediction confidence will reach its maximum value of 1. Obviously, the prediction confidence relates not only to the similarity score but also to the decision threshold.

To model the relation between the similarity score and the prediction confidence for verification problems, we first define a confidence function \(\psi(s, \tau)\) based on the similarity \(s\) and threshold \(\tau\):

\[
\psi(s, \tau) = \begin{cases} 
s - \tau, & g(s, \tau) = 1 \\
\frac{1 - s}{1 - \tau}, & g(s, \tau) = -1 
\end{cases} \quad (3)
\]

where \(s \in [-1, 1]\), and \(\tau \in (-1, 1)\). The similarity threshold \(\tau\) divides the cosine similarity interval \([-1, 1]\) into positive and negative parts, with size of \(1 - \tau\) and \(1 + \tau\), respectively, as shown in Fig. 2.

Note that face or kinship verification task is a binary decision problem. Hence, the probabilistic confidence \(c\) for a prediction \(g(s, \tau)\) based on Eq. (2) should be greater than or equal to 0.5. Then we apply an affine transformation to \(\psi(s, \tau)\) to obtain \(c(s, \tau)\):

\[
c(s, \tau) = \frac{1}{2} \psi(s, \tau) + \frac{1}{2} \quad (4)
\]

As shown in Fig. 2, if \(s\) takes a value further away from \(\tau\) (e.g., \(s = s_2\)), the prediction confidence \(c\) becomes much higher (e.g., \(c = c_2\)). On the contrary, if the value of \(s\) is closer to \(\tau\) (e.g., \(s = s_1\)), the confidence \(c\) decreases (e.g., \(c = c_1\)), indicating the more likely the model makes incorrect predictions. So far, we have proposed a flexible confidence measure that can be directly applied to any off-the-shelf verification models to yield a probabilistic estimation of prediction confidence.
**C. Confidence Calibration via Angular Scaling**

In reality, a bias exists between the verification accuracy and prediction confidence in face and kinship verification problems. This can be attributed to the fact that modern DNNs are often miscalibrated [41]. As such, the proposed confidence measure may not produce a well-matched confidence to the expected accuracy of the model. For classification tasks, it is typical to calibrate the model by scaling the logit [41]. However, if a similar process is applied to directly scale the similarity s in the verification task, the similarity s, the threshold τ, and the similarity interval will be equally scaled. This results in no change in the prediction confidence, according to Eq. (3). Inspired by the work of [12], [14], and [13], we propose a post-calibration method via angular scaling, which effectively calibrates the prediction confidence without the need for retraining the model.

We propose to calibrate the prediction confidence by calibrating the angle ϑ between pairs of faces on a recalibration dataset \( D_c = \{(X_i, Y_i), Z_i\}_{i=1}^N \), as well as a feature encoder denoted as \( f(\cdot) \). For a sample pair \((X_i, Y_i)\) with label \( Z_i \in \{-1, 1\} \) in \( D_c \), its feature representation and cosine similarity are denoted by \( x_i = f(X_i), y_i = f(Y_i) \), and \( s_i = s(x_i, y_i) \), respectively. Our Angular Scaling Calibration (ASC) learns to adjust the angle between the face pairs based on the similarity set \( S_c = \{s_i\}_{i=1}^N \) and its labels \( Z_c = \{Z_i\}_{i=1}^N \), derived from the recalibration set \( D_c \). We define our calibration function as follows:

\[
\psi(s) = \cos(\arccos(s) + w + b)
\]

where \( w \) (\( w > 0 \)) and \( b \) are learnable scalar parameters. The calibrated similarity and threshold are given by \( s' = \psi(s) \) and \( \tau' = \psi(\tau) \), respectively. The calibrated angle \( \tau' \) satisfies \( \tau' \in [0, \pi] \), where \( \tau' = \arccos(s) + w + b \).

To learn the calibration parameter \( w \) and \( b \), we minimize the following objective function:

\[
\mathcal{L}_{ASC} = \frac{1}{N} \sum_{i=1}^{N} (\psi(s_i) - Z_i)^2
\]

where \( Z_i \) equals 1 if the pair \((X_i, Y_i)\) is a positive pair, and -1 otherwise.

**Algorithm 1** Angular Scaling Calibration (ASC)

**Input:** Recalibration set \( D_c = \{(X_i, Y_i), Z_i\}_{i=1}^N \).

**Output:** Calibration parameters \( w, b \).

Set \( w = 1, b = 0 \);

Compute the similarity set \( S_c \) of feature pairs

\[
S_c = \{s(x_i, y_i)\}_{i=1}^N
\]

while not converged do

1. Update the similarity set according to Eq. (5);
2. Update \( w, b \) by optimizing the objective (6);

end

return \( w, b \)

The detail of the calibration procedure via angular scaling is summarized in Algorithm 1. In the prediction stage, we use the calibrated similarity \( s' \) and the calibrated threshold \( \tau' \) to compute calibrated prediction confidence, in accordance with Eq. (4).

Now, we briefly explain how our ASC ensures preservation of prediction accuracy. Firstly, let \( \tau \) denote the uncalibrated threshold, and we are given the positive pair \((X_p, Y_p)\) and negative pair \((X_n, Y_n)\), and their features are \((x_p, y_p)\) and \((x_n, y_n)\), respectively. It is clearly that \( s(x_n, y_n) < \tau \leq s(x_p, y_p) \), and since \( \arccos \) decreases monotonically in the interval \([-1, 1]\), we ascertain that \( \theta_n > \theta_p \geq \theta_p \). We then use \( w, b \) to adjust the angle between the two features. Considering that \( w > 0 \), if follows that \( \theta'_n > \theta'_p \geq \theta'_p \), with \( \theta'_n, \theta'_p \) are falling within \([0, \pi]\). Likewise, we can also conclude \( \cos(\theta'_n) < \cos(\theta'_p) \leq \cos(\theta'_p) \). Hence, it can be concluded that \( s'(x_n, y_n) < \tau' \leq s'(x_p, y_p) \). This indicates that our proposed ASC learns to calibrate the prediction confidence while maintaining the verification accuracy.

To evaluate the calibration performance of our proposed method for face and kinship verification, we employ the ECE (Expected Calibration Error) [51] as our evaluation metric. The ECE is computed as the weighted average difference across bins between the expected accuracy and the prediction confidence, as illustrated in Eq. (7):

\[
ECE = \frac{1}{|B_m|} \sum_{m=1}^{|B_m|} \left| \frac{\text{acc}(B_m) - \text{conf}(B_m)}{|B_m|} \right|
\]

where \( B_m \) represents all predictions fall in the \( m \)th bin, \( M \) denotes the number of bins for partition of the confidence interval \([0.5, 1]\), \( n \) is the total number of samples, \( \text{acc}(B_m) \) and \( \text{conf}(B_m) \) are the accuracy and confidence of \( m \)th bin, respectively, as calculated by Eq. (8) and Eq. (9):

\[
\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(g(s_i, \tau) = Z_i)
\]

\[
\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} c(s_i, \tau)
\]

We use equal-size interval partition [52], that is, the size of each bin is equal to \( 1/2M \). Note that an ideal calibration typically exhibits a minimal ECE. Specifically, when the accuracy of each bin is equal to its confidence, the ECE of the model will be zero, indicating that it is perfectly calibrated.

**D. Uncertainty Estimation for Confidence Calibration**

Due to our proposed confidence measure and confidence calibration methods operating at the similarity score level, they do not refer to information about image quality. For instance, in a scenario where two images are markedly blurry, thereby lacking sufficient facial information, the system may still yield high prediction confidence for these images due to elevated similarity scores. This occurrence may introduce potential risks to decision-making systems. To address this issue, we integrate uncertainty estimation into our confidence measure and confidence calibration methods. This enables the integration of image quality estimation into prediction confidence while simultaneously maintaining the calibration of confidence.
We follow the approach of [5], employing probabilistic face embedding instead of point embedding. For a given face image \( X \), we have a probabilistic embedding \( N(\mu_x, \sigma_x^2) \), where \( \mu_x, \sigma_x^2 \in \mathbb{R}^k \) correspondingly signify the mean and variance of various dimensions in the feature, and \( L \) denotes the number of feature dimension. The center \( \mu_x \) signifies the original image embedding and \( \sigma_x^2 \) is predicted via an additional network branch. To simplify computation, we assume that the original feature embedding is \( L^2 \) normalized and all dimensions are independent and identically distributed. The log likelihood of a face image pair \( (X, Y) \) being attributed to the same individual (or having a kinship) is denoted as:

\[
\log p(X = Y) = -\frac{1}{2} \sum_{l=1}^{L} \left( \frac{\mu_{x_l} - \mu_{y_l}}{\sigma_{x_l}^2 - \sigma_{y_l}^2} \right)^2 + \log(\sigma_{x_l}^2 - \sigma_{y_l}^2)
\]

where \( \text{const} = \frac{L}{2} \log 2\pi \), and \( \mu_{x_l}, \sigma_{x_l} \) refer to the \( l \)th dimension of \( \mu_x \) and \( \sigma_x \).

We maximize the log likelihood to optimize the additional network branch for accurate uncertainty estimation. The loss function is:

\[
\mathcal{L}_{PFE} = -\frac{1}{|P|} \sum_{(X_i, Y_i) \in \mathcal{P}} \log p(X_i = Y_i)
\]

where \( \mathcal{P} \) is the set of all positive pairs in the training dataset. For a given pair of face images \( (X, Y) \), with corresponding face embeddings \( (x, y) \), and the learned data uncertainty \( (\sigma_x^2, \sigma_y^2) \), the uncertainty of calibrated confidence can be derived by utilizing the uncertainty propagation [53]. This formulation is given as:

\[
\sigma_c^2(x, y, \sigma_x^2, \sigma_y^2) = \begin{cases} 
\frac{\sigma_x^2(x, y, \sigma_x^2, \sigma_y^2)}{2[1 - \psi(\tau)^2]^2}, & s(x, y) \geq \tau \\
\frac{\sigma_y^2(x, y, \sigma_x^2, \sigma_y^2)}{4[1 + \psi(\tau)^2]^2}, & s(x, y) < \tau 
\end{cases}
\]

(10)

where

\[
\sigma_x^2(x, y, \sigma_x^2, \sigma_y^2) = \frac{2}{1-s^2(x, y)} \sum_{l=1}^{L} (\psi(x, y))^2
\]

(12)

where \( \psi(\cdot) \) denotes the calibration function defined in Eq. (5), \( s(x, y) \) represents the cosine similarity function, which can be simplified as \( s(x, y) = x \cdot y \) when \( ||x||, ||y|| = 1 \). Specifically, when the calibration parameters \( w = 1 \) and \( b = 0 \), the uncertainty of the uncalibrated confidence aligns with Eq. (12). Detailed derivations pertaining to Eq. (12) and (13) are available in the Supplementary Appendix.

By introducing uncertainty estimation, the prediction confidence, denoted as \( C_c \), adheres to the normal distribution \( N(\mu_c, \sigma_c^2) \). The center \( \mu_c \) is obtained through the deterministic feature embedding following Eq. (4), which reflects the confidence of the verification process itself. And the variance \( \sigma_c^2 \) reflects the influence of the image quality on the prediction confidence. Image pairs of inferior quality exhibit a larger variance \( \sigma_c^2 \), while image pairs of superior quality demonstrate a smaller variance \( \sigma_c^2 \).

Intuitively, we can use \( \tilde{C} = \mu_c - \sigma_c \) as a substitute for \( \mu_c \) to introduce uncertainty estimation in prediction confidence, which takes into account both image quality and the image similarity. However, \( \tilde{C} \) tends to lead to an under-confident prediction confidence, thereby resulting in miscalibration. Therefore, we propose a calibrated \( \tilde{C} \) as follows:

\[
\tilde{C} = \mu_c - \alpha(\sigma_c - E(\sigma_c))
\]

(14)

where \( E(\sigma_c) \) represents the mean of \( \sigma_c \) for all pairs in the recalibration set, which reflects the average image quality. The hyperparameter \( \alpha (\alpha \geq 0) \) is the uncertainty weighting factor, and a larger \( \alpha \) indicates that the prediction confidence places greater emphasis on the uncertainty estimation. Specifically, when \( \alpha \) equals 0, the uncertainty estimation for the image is disregarded. Eq. (14) ensures that the \( \tilde{C} \) fluctuates around \( \mu_c \) and taking \( E(\mu_c) \) as its mean value, thereby guaranteeing calibration performance while incorporating uncertainty estimation.

IV. EXPERIMENTS

To validate the efficacy of our proposed methods for face and kinship verification, we conduct extensive experiments on four face and kinship datasets, including FIW [54], KinFaceW [17], LFW [55], and IJB-C [56]. In this section, we will report the accuracy, mean confidence, the ECE with \( M \) equals 20 before and after calibration, as well as the performance of uncertainty estimation on these datasets. The experimental analysis will also be presented in detail.

A. Datasets and Experimental Settings

1) FIW [54]: FIW (Families In the Wild) is the largest visual kinship recognition dataset to date, consisting of over 13,000 facial images from 1000 families, with 11 pairwise kin relationships that can be divided into three sub-groups: siblings type (i.e., sister-sister, brother-brother, and sister-brother), parent-child type (i.e., father-son, father-daughter, mother-son, and mother-daughter), and grandparent-grandchild type (i.e., grandfather-grandson, grandfather-granddaughter, grandmother-grandson, and grandmother-granddaughter). The FIW is a challenging dataset because all the images are captured in the wild with partial occlusion and large variations in background, pose, expression, illumination, and so on.

InfoNCE [57], Triplet [9], Softmax [7], and ArcFace [14] are four widely-used loss functions in face and kinship verification tasks. We first use the four loss functions to train models in FIW [54] with different backbones pre-trained on the MS1MV2 [14]. We follow the data partitioning and evaluation protocol of RFIW 2021 [58]. Moreover, we employ RetinaFace [59] for face alignment and the images are cropped into 112 \times 112 pixels.

2) KinFaceW [17]: The KinFaceW-I and KinFaceW-II datasets are widely used kinship datasets comprised of Internet-collected images of public figures and celebrities.
The two datasets contain four kin relations: Father-Son, Father-Daughter, Mother-Son, and Mother-Daughter. For these four kin relations, KinFaceW-I has 156, 134, 116, and 127 pairs of images with kin relationship. In KinFaceW-II, there are 250 pairs of images for each relationship. KinFaceW-I differs from KinFaceW-II in that, in most cases, the face pairs in the former dataset are from distinct photos while those in the latter one are from the same photos. Each image in the two datasets is aligned and resized into $64 \times 64$ for feature extraction.

We evaluate two traditional and one deep-learning kinship verification methods on the KinFaceW datasets, including: 1) NRML (LBP) [17]: for each face image, we extract 3776-dimensional uniform pattern LBP features to kinship verification. 2) NRML (HOG)[17]: first, we partition each image into non-overlapping blocks. Then, we extract a 9-dimensional HOG feature for each block and concatenate them to generate a 2880-dimensional feature. 3) InfoNCE (ResNet18)[57]: we use InfoNCE loss [57] to fine-tune a ResNet18 model pretrained on the MS1MV2 [14]. In this experiment, we use a five-fold cross-validation strategy in an image-unrestricted setting.

3) LFW [55]: LFW contains 13,233 face images collected from 5,749 individuals, of which 1,680 have 2 or more face images. We use CASIA-WebFace [60] as the training set to train different backbones with the ArcFace loss [14], and evaluate the verification performance of these models on the LFW dataset. We calibrate the confidence of different backbones using ASC in the 10-fold cross-validation scheme.

4) IJB-C [56]: IJB-C (IARPA Janus Benchmark-C) dataset contains 31,334 images and 11,779 videos captured under unconstrained condition from 3,531 subjects. There are a total of 23,124 templates, with 19,557 positive matches and 15,638,932 negative matches to enable performance evaluations at low FAR.

In this experiment, we adopt the 1:1 Mixed Verification protocol [56], that is, a single feature vector is created by taking a weighted average of the frames in each template, so that all frames in a video belonging to the same subject have the same cumulative weight as a single still image. We perform face verification on the IJB-C dataset using different backbones trained on the MS1MV2 [14] with the ArcFace loss [14] and the Triplet loss [9]. To perform confidence estimation and calibration, we use five-fold cross-validation with stratified sampling of positive and negative samples, with four of the folds being the recalibration set and the remaining one being the test set.

### B. Experimental Results on FIW

In Table I, we show the verification accuracy, the weighted average accuracy, and the ECE on FIW dataset [54] for different models and loss functions. From these results, we make the observations as below:

1) For all four loss functions, models with deeper architecture exhibit superior feature representation abilities. However, we also observe that the increase in depth leads to a negative influence on model calibration.

2) Compared with Softmax [7] and ArcFace [14] losses, InfoNCE [57] and Triplet [9] losses achieve certain performance enhancements, however, it also comes with an increase in calibration error.

3) ArcFace loss [14] leverages the angular margin penalty to enforce extra intra-class compactness and inter-class discrepancy, achieving superior verification performance with larger calibration error than Softmax loss [7].

In Fig. 3, we plot the calibration-accuracy during the training process of ResNet34 and ResNet101, intuitively showing the trend of the model’s ECE and verification accuracy with four different loss functions. It can be seen that the ECE of...
both models (ResNet34 and ResNet101) generally increase with the verification accuracy of the models. Also, if the model employs a loss directly optimizing the feature representation, such as InfoNCE [57] or Triplet [9], the model exhibits a larger ECE with the same accuracy.

In addition, we conduct experiments on the FIW dataset [54] to evaluate the effects of the ASC. Table II shows the ECE and the threshold values before and after calibration. Specifically, we adopt LBFGS [61] as our calibration optimizer with a learning rate of 0.01 and setting an upper limit of 1000 iterations. The calibrated threshold and calibration parameters (w and b) are used to evaluate the ECE on the test data.

From Table II, we observe that our ASC achieves good calibration on different models (loss functions). Additionally, it is noted that the calibrated threshold tends to converge towards 0, contributing to an equitable distribution of positive and negative pairings within the cosine similarity interval.

Fig. 4 plots the cosine similarity distribution for different models with InfoNCE loss [57] before and after calibration on the FIW dataset [54]. We observe that the similarity before calibration is condensed within a smaller interval, leading to suboptimal calibration, whereas the similarity after calibration is distributed uniformly across different similarity levels.

Fig. 5 presents accuracy-confidence plots corresponding to different models with Triplet loss [9] and Softmax loss [7] on the FIW dataset [54] before and after calibration. The plots underscore the proficiency of the prediction confidence calibration via ASC applied to all models, where ideal calibration aligns with the line $y = x$. Evidently, our ASC proves effective in achieving well-calibrated predictions, as the recalibrated confidence accurately mirrors the prediction accuracy.

Table III summarizes the average verification accuracy, average confidence before and after calibration of different models on three sub-groups of FIW [54]. It can be seen that the confidence before calibration is significantly lower or higher than the actual verification accuracy, while the recalibrated confidence aligns more closely with the ground truth. Therefore, verification models calibrated by ASC can provide a reliable confidence estimation to support decision-making in face and kinship verification systems.

C. Experimental Results on KinFaceW

In Table IV, we report the accuracy, the mean confidence, and the ECE before and after calibration on KinFaceW-I and KinFaceW-II datasets [17]. Our findings reveal that kinship verification methods concerning same-gender relationships (FS, MD) outperform those of different-gender relationships (FD, MS) in terms of verification accuracy and confidence. However, the increment in accuracy proves to be more pronounced. For NRML [17], the HOG exhibits superior verification accuracy and confidence relative to LBP, while it has a larger ECE. Compared to traditional verification methods, deep-learning methods demonstrate less confidence, with the lowest confidence and the highest ECE on most kin relationships. Besides, the accuracy is randomly distributed over the bins due to the small size of the KinFaceW dataset [17] and limited number of data samples in each bin. However, the methods calibrated by our ASC consistently show
improved calibration performance in terms of both ECE and mean ECE.

D. Experimental Results on LFW

To further evaluate our methods on face verification task, we also conduct experiments on the LFW dataset [55]. As can be seen from Table V, deeper architectures may achieve better verification accuracy, but the calibration performance gets worse, which has also been demonstrated on both FIW [54] and LFW [55]. Moreover, Fig. 6 plots the correlation between the confidence and accuracy of the four models before and after calibration. We observe that ASC is helpful to improve the calibration of all four models. Due to the high accuracy of ResNet34, ResNet50, and ResNet101 on the LFW dataset [55], the ASC is inclined to map the prediction confidence to a high-level, consequently aligning with the expected accuracy. However, as shown in Fig. 6, the calibrated confidence of all four models does not degenerate to a constant solution. Instead, it is effectively adjusted to match the prediction accuracy at varied levels.

E. Experimental Results on IJB-C

We present in Table VII the threshold $\tau$, the TAR (True Accept Rate), the accuracy, the mean confidence, and the ECE before calibration on IJB-C dataset [56]. The threshold increases as the FAR (False Accept Rate) drops from 1.0% to 0.01%, resulting in lower TAR and ECE. Upon evaluating results from four distinct groups of experiments, a consistent reduction in ECE is observed as the TAR decreases. This finding is consistent with the correlation between ECE and accuracy. Moreover, we further discover that deeper models, in the process of achieving a superior TAR, also exhibit a larger ECE. This finding aligns with observations drawn from preceding experiments, highlighting a interaction between model depth, TAR, and ECE. In addition, a focused analysis evaluating the impact of loss functions on ECE revealed significant distinctions. Specifically, the ResNet50 model, when trained using the Triplet loss [9], demonstrated a higher ECE in
TABLE VII

| Loss     | Model | FAR=1.0% T | TAR (↑) | Acc. (↑) | Conf. | ECE (↓) | FAR=0.01% T | TAR (↑) | Acc. (↑) | Conf. | ECE (↓) | AUC (↑) |
|----------|-------|-------------|---------|---------|-------|---------|-------------|---------|---------|-------|---------|---------|
| ArcFace  | ResNet101 | 0.03       | 97.17   | 99.92   | 63.54 | 34.88   | 0.22       | 93.92   | 99.94   | 70.76 | 29.64   | 99.25   |
|          | ResNet50  | -0.17      | 96.69   | 99.92   | 68.62 | 31.36   | 0.09       | 90.16   | 99.93   | 75.47 | 23.97   | 99.25   |
|          | ResNet34  | -0.20      | 96.02   | 99.89   | 71.99 | 27.98   | 0.02       | 90.56   | 99.93   | 77.56 | 21.83   | 99.74   |
|          | Triplet   | 0.15       | 93.35   | 99.71   | 64.73 | 35.10   | 0.28       | 86.31   | 99.90   | 68.31 | 31.36   | 99.07   |

Fig. 7. The trend of uncertainty vs. calibrated confidence on FIW [54] and IJB-C [56].

comparison to the model trained with the ArcFace loss [14]. Interestingly, it even exceeded the ECE of the more deeper ResNet101 model. This observation underscores the critical role that the selection of a loss function plays in affecting a model’s calibration performance.

Table VI presents the accuracy, the mean confidence, and the ECE with ASC on the IJB-C dataset [56]. Upon calibration, it is evident that the confidence closely mirrors the true accuracy and the ECE tends towards zero. These findings validate the effectiveness of our proposed ASC method, exhibiting good calibration performance across a range of model back-bones and diverse loss functions. Significantly, the ASC maintains a minimized calibration error, demonstrating its stability even when the threshold values undergo adjustments.

Table VIII

| Loss     | Model | Acc. (↑) | Conf. | Conf. w/ ASC | ECE (↓) | ECR (↓) w/ ASC |
|----------|-------|----------|-------|--------------|---------|-----------------|
| ArcFace  | ResNet50 | 74.95    | 58.59 | 70.81        | 16.36   | 5.18            |
|          | ResNet34 | 72.65    | 62.30 | 70.80        | 10.35   | 4.32            |
|          | VGG16   | 61.45    | 66.59 | 62.72        | 7.66    | 3.05            |
| InfoNCE  | ResNet50 | 75.80    | 59.57 | 74.54        | 16.23   | 4.48            |
|          | ResNet34 | 73.30    | 60.61 | 70.87        | 12.69   | 3.51            |
|          | VGG16   | 64.95    | 62.57 | 65.72        | 5.45    | 2.71            |
| ArcFace  | ResNet101 | 98.87    | 70.16 | 99.08        | 28.71   | 1.17            |
|          | ResNet50 | 98.83    | 68.36 | 99.39        | 28.45   | 1.05            |
|          | ResNet34 | 98.78    | 69.44 | 99.32        | 28.33   | 0.96            |
|          | VGG16   | 85.82    | 65.92 | 85.87        | 24.16   | 1.85            |

Fig. 8. Error vs. Reject curves showcasing the incorporation of uncertainty estimation in prediction confidence on FIW [54] and IJB-C [56].

F. Experimental Results on Cross-Dataset Verification

To further assess the generalization capability of ASC, we perform face and kinship verification under cross-dataset settings:

1) FIW(KFW-II): We calibrate the model on the FIW dataset [54], and then conduct tests on the KinFaceW-II dataset [17].

2) IJB-C(LFW): We use the IJB-C dataset [56] for calibration, and then evaluate its performance on the LFW dataset [55].

The experimental results of these cross-dataset verification are shown in Table VII. In this setting, the accuracy of all models decreases, and the prediction confidences before calibration are generally lower than the prediction accuracy, resulting in a significant ECE. However, after calibration, the prediction confidence more closely aligns with the prediction accuracy, thereby achieving ECE at a lower level. These results underscore the generalization ability and potential of ASC in cross-dataset verification settings, highlighting the feasibility of employing ASC in real-world task scenarios.

G. Experimental Results on Uncertainty Estimation

To validate the effectiveness of our proposed uncertainty estimation method, we conduct experiments on the FIW [54] and the IJB-C [56] datasets. An additional network branch is incorporated into a ResNet101 model trained using the ArcFace loss [14], with the objective of learning to predict confidence uncertainty \( \sigma_c \) as determined by Eq. (11). To facilitate comparison, the backbone was frozen during the training process.

Fig. 7 illustrates the relationship between uncertainty \( \sigma_c \) and calibrated confidence. From a statistical viewpoint, we note a decrease in uncertainty as prediction confidence heightens, suggesting an overall enhancement in image quality. To compare and analyze the impact of uncertainty estimation on decision-making, we set up the following prediction confidence schemes:

- \( C_0 \): \( C_0^w \) (without ASC)
- \( \tilde{C}_0 \): \( \tilde{C}_0^w \) (with ASC)
- \( C_1 \): \( C_1^w \) (without ASC)
- \( \tilde{C}_1 \): \( \tilde{C}_1^w \) (with ASC)
- \( C_2 \): \( C_2^w \) (without ASC)
- \( \tilde{C}_2 \): \( \tilde{C}_2^w \) (with ASC)
- \( C_3 \): \( C_3^w \) (without ASC)
- \( \tilde{C}_3 \): \( \tilde{C}_3^w \) (with ASC)

In these notations, the subscript \( w \) signifies the utilization of ASC, while \( w \)
TABLE X
COMPARISON OF ECE (%) FOR DIFFERENT POST-CALIBRATION METHODS ON FOUR DATASETS.
IN EACH GROUP, THE BEST RESULT IS MARKED IN RED

| Dataset | Loss (Method) | Model (Feature) | Uncalibrated | Histogram binning [62] | Isotonic regression [63] | ASC |
|---------|---------------|-----------------|--------------|------------------------|-------------------------|-----|
| FIW     | InfoNCE       | ResNet101       | 22.85        | 2.47                   | 2.51                    | 2.08 |
|         | ResNet50      | 16.51           | **1.57**     | 1.62                   | 1.81                    |
|         | ResNet34      | 12.72           | 2.89         | 2.60                   | 2.11                    |
|         | VGG16         | 5.97            | 2.14         | 2.17                   | 1.39                    |
|         | Triplet       | ResNet101       | 20.93        | **1.22**               | 1.36                    | 2.83 |
|         | ResNet50      | 16.97           | 2.83         | **2.82**               | 2.83                    |
|         | ResNet34      | 10.46           | 2.80         | 2.57                   | 1.87                    |
|         | VGG16         | 7.87            | **1.51**     | 4.70                   | 1.54                    |
|         | ArcFace       | ResNet101       | 19.39        | 3.14                   | 2.95                    | 2.62 |
|         | ResNet50      | 16.00           | 4.75         | 4.78                   | 2.45                    |
|         | ResNet34      | 10.04           | 3.45         | 3.76                   | 3.40                    |
|         | VGG16         | 7.69            | 1.66         | 1.40                   | 2.90                    |
|         | Softmax       | ResNet101       | 16.24        | 1.37                   | 1.20                    | 2.79 |
|         | ResNet50      | 15.77           | 2.92         | 2.85                   | 2.80                    |
|         | ResNet34      | 8.90            | 3.82         | 3.63                   | 3.56                    |
|         | VGG16         | 5.40            | 2.05         | 1.84                   | 1.81                    |
| KinFace-W-I | NRML        | LBP             | 20.95        | **1.90**               | 3.06                    | 2.14 |
|         | ResNet18      | 22.30           | 4.98         | 4.56                   | 4.39                    |
| KinFace-W-II | NRML       | LBP             | 19.38        | **1.08**               | 1.34                    | 1.53 |
|         | ResNet18      | 22.73           | 2.22         | 1.37                   | 1.62                    |
|         | InfoNCE       | ResNet18        | 26.56        | 2.35                   | 2.22                    | 1.00 |
| LFW     | ArcFace       | ResNet101       | 28.48        | **1.22**               | 0.16                    | 0.78 |
|         | ResNet50      | 27.84           | 0.98         | 0.70                   | 0.93                    |
|         | ResNet34      | 25.43           | 2.35         | 1.76                   | 1.15                    |
|         | VGG16         | 24.67           | 2.56         | 1.58                   | 1.54                    |
| IBJ-C   | ArcFace       | ResNet101       | 29.64        | 1.18                   | 0.97                    | 0.47 |
|         | ResNet50      | 23.97           | 1.49         | 1.00                   | 0.53                    |
|         | ResNet34      | 21.93           | 1.72         | 1.14                   | 0.63                    |
|         | Triplet       | ResNet50        | 31.36        | 2.24                   | **2.07**               | 2.19 |

indicates without the use of ASC; the superscript represents the value of the uncertainty weighting factor $\alpha$ in Eq. (14).

Fig. 8 displays the Error vs. Reject curves [64] for the five aforementioned schemes, in which the error rate progressively diminishes as more samples of low-confidence are excluded, indicating the effectiveness of our proposed confidence measure method. Moreover, compared to $\tilde{C}_0^w$, the use of calibrated confidence facilitates a more rapid reduction in error rate. The introduction of uncertainty estimation additionally contributes to a lower decision error rate, thereby demonstrating the efficacy of our proposed methods for confidence calibration and uncertainty estimation.

We compare the decision performance of our proposed methods ($\tilde{C}_0^w$ and $\tilde{C}_2^w$) with alternative methods, as shown in Fig. 9. These methods include:

1) PFE [5]: This method employs data uncertainty for the estimation of face image quality, using the inverse of the harmonic mean of the uncertainty in each face embedding dimension as the image quality metric.
2) SER-FIQ [6]: This method utilizes the Bayesian approximation of model uncertainty to evaluate face image quality, and we use the setting of SER-FIQ (same model) [6] to compare with other methods.
3) RetinaFace [59]: The confidence score derived from RetinaFace face detection is used as the face image quality metric.
4) DC [40]: This method uses model uncertainty and the uncertainty propagation formula [53] to estimate decision confidence, and we set $\alpha = 2$ following [40].

PFE [5], SER-FIQ [6], and RetinaFace [59] achieve suboptimal decision performance, as they solely focus on image quality and neglect the verification process. In contrast, our proposed confidence measure and confidence calibration methods demonstrate superior decision performance compared to DC [40]. Additionally, our method also offers a computational advantage in terms of efficiency.
efficiency without the need for expensive Monte Carlo Dropout [47].

Furthermore, we examine the calibration performance of the five schemes, as shown in Table IX. Compared to \( \widetilde{C}_0 \), the integration of uncertainty estimation brings about some degradation in calibration performance. However, the calibration error remains at a low level. For the FIW [54] and the IJB-C [56] datasets, setting \( \alpha \) equal to 2 is a optimal choice to balance the decision-making and the calibration performance.

\textbf{H. Comparison With Previous Post-Calibration Methods}

To validate the superiority of our ASC for confidence calibration in face and kinship verification, two widely adopted post-calibration methods are chosen for performance comparison: Histogram binning [62] and Isotonic regression [63]. While there are other post-calibration methods, such as Temperature scaling [41], Platt scaling [65], Bayesian binning [51], and Matrix and vector scaling, they are not directly applicable to confidence calibration for face and kinship verification tasks. For fair comparison, we maintain the same experimental setting across all compared methods.

\textbf{Histogram binning} [62] is a non-parametric calibration method. In this paper, Histogram binning firstly divide the cosine similarity score \( s \) evenly into \( M \) bins \( B_1, B_2, \ldots, B_M \), with bin boundaries \(-1 = a_1 < a_2 < \ldots < a_M < a_{M+1} = 1\), where the bin \( B_m \) refers to \((a_m, a_{m+1}]\). We assign each \( B_m \) a calibration score \( \eta_m \), which is optimized by:

\[
\arg\min_{\eta_1, \ldots, \eta_M} \sum_{m=1}^{M} \sum_{i=1}^{N} 1(a_m < s_i \leq a_{m+1})(\eta_m - Z_i)^2
\]  

(15)

where \( s_i, Z_i \) are uncalibrated similarity score and the label of the \( i \)th sample pair, respectively. Once the \( i \)th sample falls into bin \( B_m \), its calibrated similarity will be set to \( \eta_m \). Additionally, the calibrated threshold \( \tau' \) is updated with \( \eta_t \), where \( \tau \in (a_t, a_{t+1}] \).

\textbf{Isotonic regression} [63] is another commonly used non-parametric calibration method that can calibrate similarity score by learning a piecewise constant function \( I \), that is, \( s' = I(s) \), where \( I(s_i) \leq I(s_j) \) if \( s_i \leq s_j \). The goal of Isotonic regression is to minimize the square loss of \( \sum_{i=1}^{N}(I(s_i) - Z_i)^2 \), which is optimized to learn the boundaries and calibration scores of each interval. We employ PAVA [66] algorithm to optimize the function \( I \), and the calibrated threshold \( \tau' \) is set to \( I(\tau) \).

Table X shows the calibration of three alternative methods (ASC, Histogram binning [62], and Isotonic regression [63]) on the FIW [54], KinFaceW [17], LFW [55], and IJB-C [56] datasets, utilizing the ECE metric. We observe that ASC achieves the best calibration performance in most cases. ASC is easy to implement, as it involves only two parameters to be optimized and they can be efficiently learned by most gradient-based optimizers. In addition, ASC has...
a differentiable calibration function, making it seamlessly integrable with uncertainty estimation.

I. Visualization Analysis

Fig. 10 illustrates prediction confidence on sampled pairs from three datasets. These pairs are categorized into three groups based on the calibrated confidence: low confidence, medium confidence, and high confidence. The visualization results indicate that ASC can appropriately increase or decrease prediction confidence to match the true accuracy, providing accurate confidence estimation for system decision-making. In addition, samples with low confidence are usually accompanied by certain facial occlusions or image blurring. However, there are also some samples that lack facial information but have high confidence (e.g., the case in the top right). The application of uncertainty estimation enables us to adjust prediction confidence by taking into account both image quality and the image similarity, thereby bolstering the reliability and trustworthiness of the verification system.

V. CONCLUSION

In this work, we propose a simple yet effective confidence measure for face and kinship verification tasks, allowing any off-the-shelf verification models to estimate the decision confidence in an efficient and flexible way. We further introduce a confidence calibration method using angular scaling for face and kinship verification, which is retraining-free and accuracy-preserving. Incorporating our proposed confidence measure and confidence calibration methods with uncertainty estimation, we succeed in jointly model the image quality and verification process. This integration results in a notable improvement in decision-making performance while maintaining calibration performance. Comprehensive experiments are conducted on four widely used face and kinship verification datasets to examine the calibration of prevalent facial and kinship verification models. The effectiveness of our proposed methods is validated through these experiments. A comparative analysis with two popular post-calibration methods shows that our ASC exhibits superior calibration performance. In future research, we will explore improving the generalization ability and versatility of our approach, adapting it to diverse application domains such as medical imaging or the manufacturing industry.

REFERENCES

[1] S. Chopra, R. Hadsell, and Y. LeCun, “Learning a similarity metric discriminatively, with application to face verification,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2005, pp. 539–546.
[2] R. Fang, K. D. Tang, N. Snively, and T. Chen, “Towards computational models of kinship verification,” in Proc. IEEE Int. Conf. Image Process., Sep. 2010, pp. 1577–1580.
[3] R. Liu and W. Tan, “EQFace: A simple explicit quality network for face recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2021, pp. 1482–1490.
[4] W. Xie, J. Byrne, and A. Zisserman, “Inducing predictive uncertainty estimation for face recognition,” in Proc. Brit. Mach. Vis. Conf. (BMVC), 2020, pp. 1–13.
[5] Y. Shi and A. Jain, “Probabilistic face embeddings,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6901–6910.
[6] P. Terhorst, J. N. Koll, N. Damer, F. Kirchbuecher, and A. Kuiper, “SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 5650–5659.
[7] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1701–1708.
[8] Y. Sun, D. Liang, X. Wang, and X. Tang, “DeepID3: Face recognition with very deep neural networks,” 2015, arXiv:1502.00873.
[9] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 815–823.
[10] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 499–515.
[11] W. Liu, Y. Wen, Z. Yu, and M. Yang, “Large-margin softmax loss for convolutional neural networks,” in Proc. 33rd Int. Conf. Mach. Learn., 2016, pp. 507–516.
[12] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “SphereFace: Deep hypersphere embedding for face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6738–6746.
[13] H. Wang et al., “CosFace: Large margin cosine loss for deep face recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5265–5274.
[14] 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. 4685–4694.
[15] J. P. Robinson, M. Shao, and Y. Fu, “Survey on the analysis and modeling of visual kinship: A decade in the making,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 8, pp. 1577–1580, Aug. 2022.
[16] X. Wu et al., “Facial kinship verification: A comprehensive review and outlook,” Int. J. Comput. Vis., vol. 130, no. 6, pp. 1494–1525, Jun. 2022.
[17] J. Lu, X. Zhou, Y-P. Tan, Y. Shang, and J. Zhou, “Neighborhood repulsed metric learning for kinship verification,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 2, pp. 331–345, Feb. 2014.
[18] H. Yan, J. Lu, W. Deng, and X. Zhou, “Discriminative multimetric learning for kinship verification,” IEEE Trans. Inf. Forensics Security, vol. 9, no. 7, pp. 1169–1178, Jul. 2014.
[19] X. Zhou, Y. Shang, H. Yan, and G. Guo, “Ensemble similarity learning for kinship verification from facial images in the wild,” Inf. Fusion, vol. 32, pp. 40–48, Nov. 2016.
[20] J. Hu, J. Lu, L. Liu, and J. Zhou, “Multi-view geometric mean metric learning for kinship verification,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2019, pp. 1178–1182.
[21] M. Xu and Y. Shang, “Kinship measurement on face images by structured similarity fusion,” IEEE Access, vol. 4, pp. 10280–10287, 2016.
[22] X. Zhou, H. Yan, and Y. Shang, “Kinship verification from facial images by scalable similarity fusion,” Neurocomputing, vol. 197, pp. 136–142, Jul. 2016.
[23] S. Wang, J. P. Robinson, and Y. Fu, “Kinship verification on families in the wild with marginalized denoising metric learning,” in Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), May 2017, pp. 216–221.
[24] J. Lu, J. Hu, and Y. P. Tan, “Discriminative deep metric learning for face and kinship verification,” IEEE Trans. Image Process., vol. 26, no. 9, pp. 4269–4282, Sep. 2017.
[25] X. Qin, D. Liu, and D. Wang, “Social relationships classification using social contextual features and SVDD-based metric learning,” Appl. Soft Comput., vol. 77, pp. 344–355, Apr. 2019.
[26] Q. Duan, L. Zhang, and W. Zuo, “From face recognition to kinship verification: An adaptation approach,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2017, pp. 1590–1598.
[27] S. Wang, Z. Ding, and Y. Fu, “Cross-generation kinship verification with sparse discriminative metric,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 11, pp. 2783–2790, Nov. 2019.
[28] N. Kohli, D. Yadav, M. Vatsa, R. Singh, and A. Noore, “Deep face-representation learning for kinship verification,” in Deep Learning in Biometrics. CRC Press, 2018, p. 127.
[29] K. Zhang, Y. Huang, C. Song, H. Wu, and L. Wang, “Kinship verification with deep convolutional neural networks,” in Proc. Brit. Mach. Vis. Conf., 2015, pp. 1–12.
[30] X. Zhou, K. Jin, M. Xu, and G. Guo, “Learning deep compact similarity metric for kinship verification from face images,” Inf. Fusion, vol. 48, pp. 84–94, Aug. 2019.
