Audio-Visual Kinship Verification: A New Dataset and a Unified Adaptive Adversarial Multimodal Learning Approach

Xiaoting Wu, *Member, IEEE*, Xueyi Zhang, Xiaoyi Feng, Miguel Bordallo López, and Li Liu, *Senior Member, IEEE*

***Abstract*—Facial kinship verification refers to automatically determining whether two people have a kin relation from their faces. It has become a popular research topic due to potential practical applications. Over the past decade, many efforts have been devoted to improving the verification performance from human faces only while lacking other biometric information, for example, speaking voice. In this article, to interpret and benefit from multiple modalities, we propose for the first time to combine human faces and voices to verify kinship, which we refer it as the *audio-visual kinship verification* study. We first establish a comprehensive audio-visual kinship dataset that consists of familial talking facial videos under various scenarios, called TALKIN-Family. Based on the dataset, we present the extensive evaluation of kinship verification from faces and voices. In particular, we propose a deep-learning-based fusion method, called unified adaptive adversarial multimodal learning (UAAML). It consists of the adversarial network and the attention module on the basis of unified multimodal features. Experiments show that audio (voice) information is complementary to facial features and useful for the kinship verification problem. Furthermore, the proposed fusion method outperforms baseline methods. In addition, we also evaluate the human verification ability on a subset of TALKIN-Family. It indicates that humans have higher accuracy when they have access to both faces and voices. The machine-learning methods could effectively and efficiently outperform the human ability. Finally, we include the future work and research opportunities with the TALKIN-Family dataset.**

***Index Terms*—Adversarial learning, audio-visual, benchmark dataset, kinship verification, multimodal fusion.**

Manuscript received 18 May 2022; revised 30 August 2022; accepted 26 October 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB3100800; in part by the Academy of Finland under Grant 331883 and Grant 346208-6G Flagship; in part by the National Natural Science Foundation of China under Grant 61872379; in part by the University of Oulu and the Academy of Finland under Grant Profi6 336449; in part by the Key Research and Development Program of Shaanxi under Grant 2022ZDLGY06-07; and in part by the Shenzhen Science and Technology Program under Grant GJHZ20200731095204013. This article was recommended by Associate Editor J.-H. Xue. *(Corresponding author: Li Liu.)*

Xiaoting Wu is with the Center for Machine Vision and Signal Analysis, University of Oulu, 90570 Oulu, Finland, and also with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710060, China.

Xueyi Zhang is with the College of System Engineering, National University of Defense Technology, Changsha 410073, Hunan, China.

Xiaoyi Feng with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710060, China, and also with the Research and Development Institute, Northwestern Polytechnical University (Shenzhen), Shenzhen 518063, China.

Miguel Bordallo López is with the Center for Machine Vision and Signal Analysis, University of Oulu, 90570 Oulu, Finland, and also with the Cognitive Technologies for Intelligence, VTT Technical Research Centre of Finland, 90570 Oulu, Finland.

Li Liu is with the College of System Engineering, National University of Defense Technology, Changsha 410073, Hunan, China, and also with the Center for Machine Vision and Signal Analysis, University of Oulu, 90570 Oulu, Finland (e-mail: li.liu@oulu.fi).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2022.3220040.

Digital Object Identifier 10.1109/TCYB.2022.3220040

I. Introduction

FACIAL kinship verification (FKV) aims at automatically determining whether two individuals have a kin relationship or not from their given facial images or videos [1]. Since the seminar work by Fang et. al [2], recently FKV has gained increasing attention [1] due to its wide range of potential applications, including finding missing persons, border control and customs, criminal investigations [3], family photo album organization, improving the performance of face recognition systems, and social media analysis [4]. To the best of our knowledge, although closely related to face verification, that has been well developed and made into products for real world [5], the FKV technology has not been, however, capable of performing at a sufficient level for any practical applications due to its unique challenges discussed in great detail in [1].

Existing research in kinship verification has been extensively focused on exploring kinship features from the visual modality of the facial images/videos [3], [6], [7]. Certainly, facial similarity plays an important role in FKV, as facial similarity and kinship judgments are highly correlated according to recent psychology research [8]. However, there have been studies demonstrating that voice similarity is also related to kinship judgments [9], [10], [11], [12], [13]. For instance, according to [9], the vocal tract shape that affects voice properties are genetically determined, consequently subjects with kinship have a similar voice. In addition, the study on human perception of kin voice indicates that humans have the ability to judge kinship by listening to the speaking voice [14], [15]. Despite these evidences, voice modality has not yet been explored for FKV.

In recent years, audio-visual fusion has been shown to be an effective way of improving performance in various problems, including emotion recognition [16], speech recognition [17], event detection [18], and biometrics [19], such as speaker identification and speaker authentication. Based on the aforementioned discussion, it is natural to ask: in addition to the

visual modality, is it beneficial to explore other modalities (the voice channel in specific) for the problem of FKV? Therefore, in order to answer this question, in this article, we carry out the first study, and aim to build an audio-visual kinship verification framework in an attempt to further improve the FKV performance. To this end, we need to address two main challenges: 1) collection and publication of a new audio-visual dataset as there is no such datasets available and 2) development of novel approaches specifically for improving the verification performance.

High-quality datasets enable rapid progress in the FKV task. In our previous preliminary attempt [20], we have collected the TALKIN dataset. However, the TALKIN has some obvious limitations, that is, limited number of training samples, limited diversity in terms of environmental conditions, kinship categories, and mono-annotation with *binary kinship labels* only. To address some of these identified limitations, we aim to establish a new audio-visual kinship dataset, called TALKIN-Family, that consists of facial videos and synchronous speaking audio with properties that differ from the existing one. In TALKIN-Family, there are 246 unique family trees and 1012 individuals with rich annotations of family relationships, age, gender, and scene conditions. The size of family tree ranges from 2 to 14 subjects whose age is between 5 and 81 years old. Each subject has multiple talking facial videos of about 10-s length under different conditions. Overall, there are 9.2 h of videos in TALKIN-Family.

In response to the second challenge, we consider the design of a novel framework for the task of audio-visual FKV. It encompasses two main steps, that is: 1) extracting appropriate features and 2) integrating modality information. Representing modalities, that is, *audio* and *video*, in an appropriate way is crucial before fusion. Visual features have been widely studied for FKV [21]. Comparatively, very few acoustic features are designed specially for kinship verification, because the study has been largely under explored. However, well-known acoustic representations such as Mel-frequency cepstral coefficients (MFCCs [22]) and data-driven features [23], [24] have been commonly applied in speech community. Similar to the correlation between facial similarity and FKV, we propose to compute the voice similarity and set new benchmark methods for FKV by using acoustic features.

When fusing audio-visual features for the problem of FKV, based on our benchmarks and investigation, we find the inter modal discrepancy and modal weighting are essential to exploit informative knowledge. Motivated by the adversarial learning [25] strategy and the self-attention mechanism [26], we propose the fusion method, called unified adaptive adversarial multimodal learning (UAAML) based on deep neural networks (DNNs), which addresses the aforementioned challenges. The UAAML jointly considers multimodal feature learning and kinship attention weights with similarity learning. Particularly, we introduce the L_2 norm layer [27] to generate the unified features before fusion and make the network training stable and efficient.

We highlight the contributions of this article as follows.

1) We address a new task of *audio-visual kinship verification* and investigate to exploit the human voice as the complimentary feature to solve the FKV problem. The largest and most comprehensive audio-visual kinship dataset, namely, TALKIN-Family is proposed.
2) We propose a multimodal fusion network, that is, UAAML, which can jointly learn modal invariant and the attentive features with the unified multimodal features for kinship verification.
3) The extensive benchmark evaluations are conducted on the TALKIN-Family dataset. The experimental results set the benchmark for the voice-based FKV problem and demonstrate that the vocal features provide complementary information over facial features.
4) The proposed method UAAML achieves the overall competitive performance compared with the baseline methods. A set of ablation studies and evaluation on different conditions also indicate our technical contributions with improvements in kinship verification.[1]

This is an extended version of our conference paper [20]. We improve our previous work from aspects of the proposed dataset and evaluated methods, including benchmarks and the proposed fusion method. The human performance on kinship verification from faces and voices is also studied. The structure of this article is organized as follows. Section II briefly reviews related work. Section III introduces the details of the TALKIN-Family dataset. Section IV presents our proposed UAAML approach. Section V shows extensive experiments and results. Section VI concludes this article with possible future directions.

II. Related Work

A. Kinship Verification

1) Kinship Datasets: Table I compares the main characteristics of existing kinship datasets. We categorize those datasets with data modality. At the early years, kinship datasets are mainly based on images. Among those, the FIW dataset is the largest and most comprehensive image kinship dataset. The facial video kinship datasets are the ones that only the facial information is available, including UvA-NEMO Smile [32], KFVW [33], and KIVI [34]. The video and audio kinship datasets include the TALKIN dataset [20] and FIW MM dataset [35]. The TALKIN dataset is the first audio-visual kinship dataset. It was organized with a pairwise structure, while lacking the family structure. Each subject has only one video sample under the unconstrained condition. FIW MM is the recent one and has a larger data volume with 200 families, and multiple samples were collected for some subjects under wild conditions. Compared with TALKIN and FIW MM datasets, the dataset proposed in this article, that is, TALKIN-Family, is superior on the dataset volume and environment scenarios. More families are included in the dataset. Moreover, the TALKIN-Family dataset also contains people speaking the fixed content and different contents.

2) Kinship Verification Methods: The kinship verification has been studied for more than ten years [1]. The research is mainly carried out from still facial images. Early image-based works focused on extracting the facial features with handcrafted descriptors [30], [36] and measuring the similarity by computing common distance metrics such as cosine

[1]The dataset and benchmark codes will be made available for download along with this article publication.

TABLE I
MAIN CHARACTERISTICS OF EXISTING KINSHIP DATASETS. WE SORT THOSE DATASETS BY THE DATA MODALITY. IN THE EARLY YEARS, MANY IMAGE KINSHIP DATASETS WERE PROPOSED. THEN, SOME VIDEO DATASETS WITH ALIGNED FACIAL INFORMATION WERE PROPOSED. THE DATASET PROPOSED IN THIS ARTICLE, TALKIN-FAMILY, CONSISTS OF BOTH VISUAL AND VOCAL INFORMATION AND IS THE MOST COMPREHENSIVE ONE BY FAR

Modality	Dataset	Year	Size	Family structure	Multiple sample	Controlled environment
Image	CornellKin [2]	2010	150 pairs	○	○	○
	UB Kinface [28]	2011	200 groups	○	○	○
	Family 101 [29]	2013	101 families	●	●	○
	KinFaceW-I [3]	2014	533 pairs	○	○	○
	KinFaceW-II [3]	2014	1000 pairs	○	○	○
	TSKinFace [30]	2015	1015 groups	○	○	○
	FIW [21]	2016	1000 families	●	●	○
	WVU [31]	2017	113 pairs	○	●	○
Video	UvA-NEMO Smile [32]	2012	1240 videos	○	●	●
	KFVW [33]	2018	418 pairs	○	○	○
	KIVI [34]	2019	211 families	●	○	○
Video & Audio	TALKIN [20]	2019	400 pairs	○	○	○
	FIW MM [35]	2021	200 families	●	●	○
	TALKIN-Family (this paper)	2022	246 families	●	●	● ‖ ○

similarity [36]. Then, the metric-learning-based methods [3], [37] were proposed to separate the kin and nonkin pairs. With the development of deep learning, many methods with different motivations were raised [1]. The first End-to-End deep learning architecture for kinship verification is proposed by Zhang et al. [38] in 2015. The network takes two stacked facial images as the input and then predicts the kinship at the top layer. Later, Li et al. [39] proposed a Siamese network with the similarity metric to learn the discriminative features for kinship verification. Based on the Siamese CNN architecture, different strategies were explored on how to reason the relations between two facial features. Dahan and Keller [40] computed the kinship verification scores by fusing the face embeddings collected from the last FC layer. Li et al. [41], [42] introduced the star-shaped graph to model the facial feature. Then, the relational reasoning is performed on the graph by the recursive message passing scheme. The kinship dataset has the intrinsic issue of limited positive samples and far more negative samples. To exploit all the possible training samples, Li et al. [43] proposed the meta-mining approach to sample the unbalanced training batch. Alternatively, Song and Yan [44] proposed the KinMix method to augment the kinship positive samples with the linear sampling method from the feature level. Extensive experiments showed that the refined training batch could effectively boost the model learning capability.

On the basis of image-based studies, to capture multisource information, researchers proposed to study the kinship verification from facial videos. Compared with image-based studies [21], the works on facial videos [32], [33], [34] can only be found with a limited scale [1]. At the beginning, the constrained facial video dataset was used. Dibeklioglu et al. [32] proposed to fuse the facial appearance and dynamic facial features that are extracted from a smiling video clip. However, collecting standard smiling faces under unconstrained conditions is relatively hard. Therefore, researchers raised the study of kinship verification from unconstrained facial videos. Kohli et al. [34] extracted the spatiotemporal kin information in videos. Yan and Hu [33] studied the metric-learning methods on unconstrained videos for kinship verification. However, the works above neglect the additional kinship clue that resides in the human voice.

B. Acoustical Study for Kinship

In our daily lives, people with a kin relation can have similar voices. For instance, it is sometimes hard to distinguish between father and son over the phone. This phenomenon has attracted researchers from many domains into the fields. Researchers explicitly studied the vocal similarity of kin people. The earliest genetics of voice research was found in the 1990s. Sataloff [9] demonstrated that the voice function is related to the phonatory organ structures. The physical features are genetically determined, which intuitively indicates that the human voice is also genetically determined. Later, psychological studies assessed human perception on recognizing the kin voice. Studies by Van et al. [14] and Taylor [15] showed that humans could verify kinship from voice by providing listeners with the voice of specific sentences. Motivated by the research above, acoustic studies quantitatively confirmed voice similarity within kinship by measuring and comparing various acoustic characteristics [10], [11], [13]. Though many works have been carried out on studying the vocal similarity of kinship, the voice has not been directly applied in automatic kinship verification.

C. Multimodal Learning

Multimodal fusion methods can exploit complementary sources of information. Different sources of information are typically integrated through early fusion (feature level) or late fusion (score or decision levels) [45]. Feature-level fusion using concatenation or aggregation is often considered to provide a high level of accuracy. However, feature patterns may also be in compatible and increase system complexity. Techniques for score-level fusion using deterministic (e.g., average fusion) or learned functions are commonly employed but are sensitive to the impact of score normalization methods on the overall decision boundaries.

When considering multimodal fusion, one main challenge is eliminating the modal discrepancy and learning a joint feature space that can better fuse the features. Recent generative adversarial networks (GANs) [25], [46] have achieved the significant success that can map the data distribution into the desired one by adversarial training. Inspired by this,

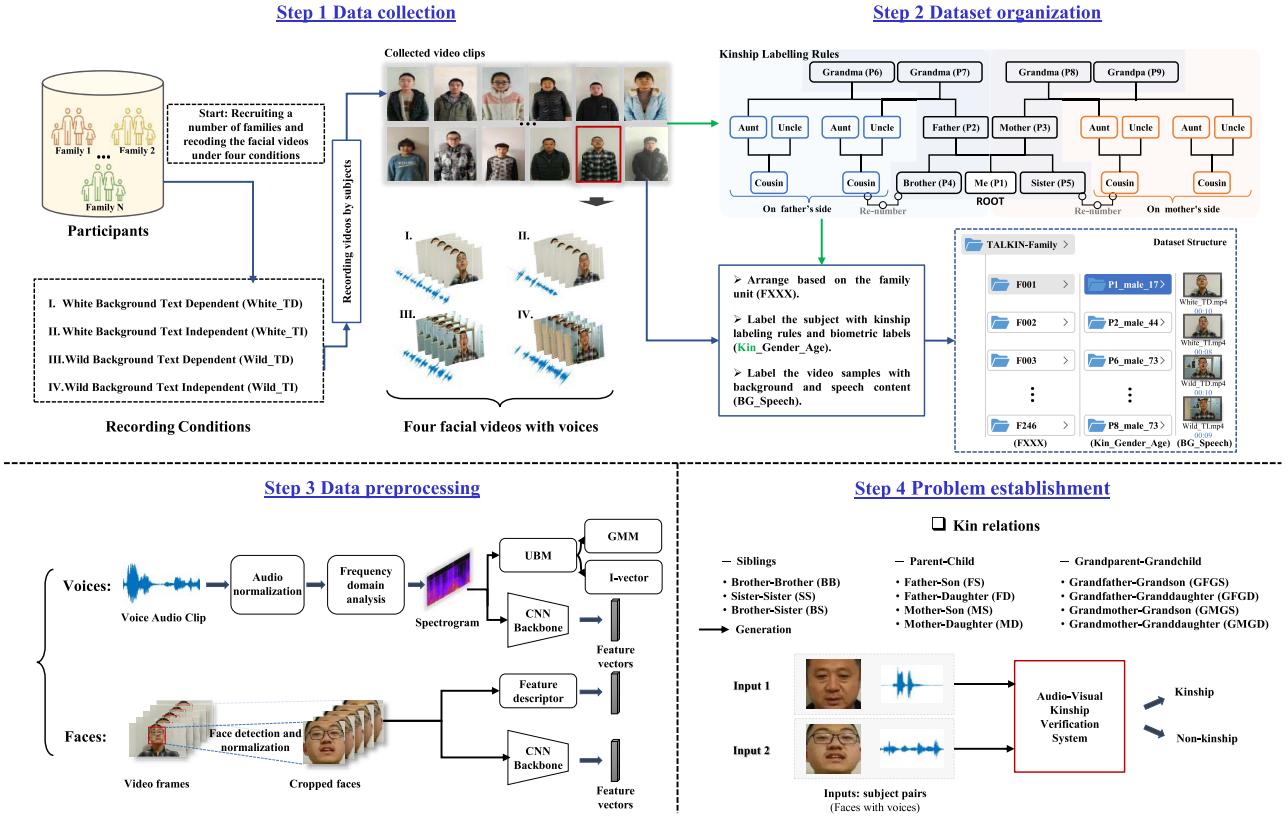

Fig. 1. Overall collection pipeline for the TALKIN-Family dataset. The data in TALKIN-family is collected offline by recruiting a number of families. Subjects participate the data collection with their family. Each subject has four facial talking videos under two background and two speech conditions. The TALKIN-Family is organized with family structure, and in each family, people are labeled according to our kinship labeling rules. Then, we do data preprocessing with audio and facial video separately. To study the audio-visual kinship verification, we define the problem with different kin relation types.

Mai et al. [47] built the encoder–decoder networks for different modal inputs to learn the latent feature embeddings. The adversarial learning was introduced on the encoder to learn the joint feature space for different modalities. Zhou and Shen [48] studied the multimodal clustering problem. They developed the end-to-end adversarial attention multimodal clustering (EAMC) method that consisted of the adversarial learning module and modal attention module to align the feature distribution and quantify the important modal weights. A proposed clustering objective was added to guide the network training on the top of the network.

III. TALKIN-Family Dataset

A. Motivations

Benchmark datasets serve as the common ground for performance measurement and comparison of various algorithms, and help the field progress toward challenging problems. On the other hand, dataset biases could bring unwanted information, which the system takes as class clues and show high confidence in prediction [49], [50]. To ensure our kinship dataset applicable, the possible familial biases, such as recording devices, recording conditions, and speech contents are considered during the data collection procedure. We found that video-sharing websites such as YouTube[2] usually contain free-style speaking videos while lacking fixed-text speech. To fill the blank, we choose to collect the TALKIN-Family offline. The video recording task is distributed to the participating families, and family members record the qualified videos by following the provided instructions. We will introduce the collection steps in details in the following section.

[2]YouTube is a popular U.S.-based video-sharing website.

B. Collection Pipeline

The overall collection pipeline is shown in Fig. 1. The participants were asked to record the frontal talking facial videos of themselves and biologically related family members. Considering eliminating the family-related biases (e.g., recording conditions, recording devices, and speech contents), we set up several recording protocols.

Participants: The subjects involved in the recording mission within one family should be biologically related. The number of subjects within one family should be at least two, including collateral relatives and direct relatives across the generations. This means that collateral relatives cannot be considered an isolated family. Subjects across different families should have no biological connection.

Environment Conditions: The background is quiet without noise or voices from other people. There is only one subject that appears in the video. To further ensure that videos within one family do not only have one background, we ask the subject to record videos against both the white background and the nonwhite one. We refer to the white background as "white" and the nonwhite background as "wild," as shown in

Fig. 1. This could eliminate the familial background bias [50] by generating kin pairs across different backgrounds.

Speech Content: In the speaker verification study, it is distinguished as *text-dependent speaker verification* and *text-independent speaker verification*. When the speaking content is fixed, it refers to text-dependent speaker verification [51]. In text-independent speaker verification, subjects talk freely without the explicit cooperation [52]. In our dataset, we consider both scenarios for the sake of extensive usage of TALKIN-Family and meanwhile avoiding familywise spoken utterances. The participants were provided with the specific content (that is the Mandarin new year greeting). Other than that, they could speak freely while differently from the provided content. The abbreviations for text-dependent and text-independent are TD and TI. Therefore, for each subject, there are four talking videos, referred as *BACKGROUND_CONTENT* (i.e., *White_TD*, *White_TI*, *Wild_TD* and *Wild_TI*), as shown in Fig. 1.

Shooting Device: The videos were recorded by the camera of the smartphone. The phone should be held still during the recording, and the retouching function was turned off. Within one family, multiple (more than one) phones were asked to be used for recording the videos (to avoid device bias). Each video lasts for about 10 s.

Data Packing: We set the principle subject as ROOT ("me"), who is one of the young generations. Family members are backtracked based on the root, and the family tree is generated and labeled as in Fig. 1. Every involved single-family has a family folder as FXXX (i.e., F001-F246). In addition, the gender and age labels were also collected. Under the family folder, each subject has a subfolder called ID_GENDER_AGE (e.g., P1_female_6), where ID refers to the subject's family role defined by the family tree. GENDER is male or female, and AGE is an integer referring to the subject's age. Then, under the subject's folder, four facial videos are stored.

C. Data Preparation

In the TALKIN-Family dataset, each video clip was recorded with the cooperation of the participants, and only one subject appears in each video. Therefore, Speaker Diarization [53] is not required to determine "who spoke when" before data preprocessing. We do the preprocessing from visual data and audio data separately, as described below.

Visual Data: We first extract facial frames from each video, and faces are automatically detected, cropped, and aligned as done in [54]. Note that some recorded videos are shot in landscape mode or upside down. Therefore, in such cases, face orientation and image rotation are needed during face detection. Then, facial frames are resized to 224×224 and encoded by face-image descriptors. Section V details the face descriptors we employed in the experiments, including traditional descriptors and deep encoders.

Audio Data: Since the subject starts to talk and ends right after the subject stops, we extract the audio directly from videos and save them as WAV files. The signal is converted and normalized to the single channel at a 44.1-kHz sample rate. Standard methods in the speech field, MFCCs [22] and DNNs, are used to embed the audio features.

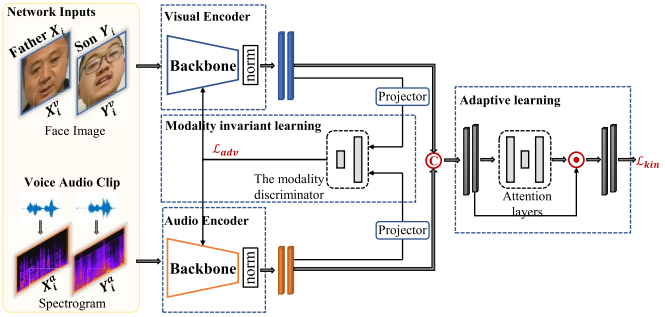

Fig. 2. Proposed UAAML method.

D. Dataset Statistics

1) Familial Information: TALKIN-Family is organized with family structure, and it contains 246 families. Each family has 2–14 family members. There are total of 1012 subjects and 4048 clips of videos in the dataset. The age of the subjects varies between 5 years and 81 years old.

2) Data Details: The length of each video clip is about 10 s. In total, TALKIN-Family has 9.2 h of videos. There are about 1 million facial frames in TALKIN-Family. All the subjects are from China and speak Mandarin Chinese (some of those have accents).

E. Problem Establishment

We address the audio-visual kinship verification as a binary classification problem: given a pair of signals [a pair of video sequences with speech utterances, for example, $(\mathbf{X}, \mathbf{Y})$], the objective is to automatically determine whether they have a kin relation. In practice, we represent $\mathbf{X}$ and $\mathbf{Y}$ using recording-level representations. The kinship score, a numerical indicator associated with higher values for kin relation pairs, is obtained by computing similarity score between the feature representations. Three levels of generation (Siblings, Parent–Child, and Grandparent–Grandchild) are considered in our experiments.

IV. Proposed Method

The overall framework of the proposed method is shown in Fig. 2. It consists of modality-specific feature generators, modal fusion, and kinship assignment. The modality-specific networks are encouraged to exploit the distinct modal property. Then, the modal fusion is trained to eliminate the cross-modal discrepancy to parse the better fusion of multiple feature vectors from different modalities. When obtaining the fused features, the contrastive loss is added to enforce the network to learn the compactness within kinship and separation between nonkinship.

A. Preliminaries

Let $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{Y}_i, l_i) | i = 1, 2, \ldots, N\}$ be the training set of N sample pairs, where $\mathbf{X}_i = \{\mathbf{X}_i^a, \mathbf{X}_i^v\}$, $\mathbf{Y}_i = \{\mathbf{Y}_i^a, \mathbf{Y}_i^v\}$. $\mathbf{X}_i$ and $\mathbf{Y}_i$ represent the ith sample pair that comes with both audio and visual modalities denoted by $\mathbf{X}_i^a, \mathbf{X}_i^v$ and $\mathbf{Y}_i^a, \mathbf{Y}_i^v$,

respectively. The pairwise label l_i denotes whether the ith the sample pair has a kin relation, that is, $l_i = 1$ represents that $\mathbf{X}_i$ and $\mathbf{Y}_i$ have a kin relation, and $l_i = 0$ denotes that $\mathbf{X}_i$ and $\mathbf{Y}_i$ have the nonkin relation.

Our method has two feature encoders: 1) the audio encoder $E_a(\cdot;\theta_a)$ and 2) visual encoder $E_v(\cdot;\theta_v)$ that are parameterized by θ_a and θ_v. The audio and visual data are fed into the modal-specific encoder, and the feature representation is expected to be modal invariant. This is achieved by the adversarial learning associated with the discriminator $D(\cdot;\theta_d)$, where θ_d is the network parameter. Besides, to let the feature pay more attention to effective kinship traits and emphasize them, the attention mechanism is proposed to learn the weights for the feature-level fusion. The weight vector $\mathbf{w}$ is computed by the multiple layer perceptron (MLP). The entire network is designed with Siamese fashion that shares weights for two different inputs $\mathbf{X}_i$ and $\mathbf{Y}_i$. To preserve the kin discrimination of the network, we employ the contrastive loss L_{kin} to let the modal learn the closeness of kinship and separation of nonkinship.

B. Modality-Specific Networks

Different sources of data are difficult to be combined at the raw data level. Therefore, we first adopt the modality-specific networks to transform the face and voice data into the latent feature space. Following the work in [20], the network inputs are the facial image and spectrogram computed from a particular speech. The residual network (ResNet) architecture [55] is adopted for both face and voice backbone network described as follows. We take sample $\mathbf{X}_i$ for an example, which goes the same to the input $\mathbf{Y}_i$.

1) Visual Subnet: The visual backbone directly adopts the InsightFace with ResNet-34 architecture [56], [57]. Given an input facial image $\mathbf{X}_i^v \in \mathbb{R}^{D\times H\times W}$, we extract the corresponding feature embedding as $\mathbf{x}_i^v = E_v(\mathbf{X}_i^v)$. The W and H indicate the spatial size, and D is the number of channels. As the facial image is cropped and resized into 112×112, the generated facial features fall into 512-D.

2) Audio Subnet: The audio backbone employs the ResNet-50 pretrained on Voxceleb2 [23], [58] to extract the vocal features from the spectrogram inputs. We extract a 3-s utterance clip and convert it into the single channel with a 16-kHz sampling rate. The spectrogram is generated by a sliding hamming window of width 25 ms and step 10 ms. Therefore, the audio network input $\mathbf{X}_i^a$ has the size of 512×300, and the corresponding output $\mathbf{x}_i^a = E_a(\mathbf{X}_i^a)$ is a 2048-D feature embedding.

Similarly, we can have the audio and visual embedding for $\mathbf{Y}_i$ as $\mathbf{y}_i^a = E_a(\mathbf{Y}_i^a)$, $\mathbf{y}_i^v = E_v(\mathbf{Y}_i^v)$.

C. Model Fusion

The modal fusion module fuses audio and visual features for comprehensive estimation. It consists of the unified feature operation, modal alignment, and feature fusion attention learning.

1) Multimodal Adversarial Learning: When merging features generated from different modalities, they generally have different scales and norms. Directly combining these features leads to poor fusion performance since the larger features can overwhelm the smaller ones. Rather than carefully tuning the network parameters with efforts, Liu [27] found that normalizing the features before fusion improves the model stability. Therefore, before learning the modal-invariant features, we add a L_2 normalization layer to transform the feature as a unified one. Formally, for the audio feature $\mathbf{x}_i^a$ and visual feature $\mathbf{x}_i^v$, we normalize them differently as $\hat{\mathbf{x}}_i^a$, $\hat{\mathbf{x}}_i^v$ using L_2-norm

$$F(\mathbf{x}) = \hat{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$$
$$\text{s.t. } \mathbf{x} = \{x_1, x_2, \ldots, x_d\},\ \|\mathbf{x}\|_2 = \left(\sum_{i=1}^{d}|x_i|^2\right)^{\frac{1}{2}}. \tag{1}$$

The audio and visual encoders learn multimodal representations that may have a large gap between different modalities. Inspired by the recent GANs [25], we introduce the discriminator $D(\cdot;\theta_d)$ to distinguish the audio and visual features. Since the audio and visual features have different dimensions, we first feed them into one fully connected layer that is $FC_a(\cdot)$ and $FC_v(\cdot)$ to map them into a common length. Then, the two-class classification is performed. The discriminator is optimized by the following objective function:

$$\min_{\theta_d} \mathcal{L}_d = -\mathbb{E}_{\mathbf{X},\mathbf{Y}\in\mathcal{D}} \sum_{i=1}^{N} \log\big(D(\hat{\mathbf{x}}_i^a)\big) + \log\big(1 - D(\hat{\mathbf{x}}_i^v)\big) + \log\big(D(\hat{\mathbf{y}}_i^a)\big) + \log\big(1 - D(\hat{\mathbf{y}}_i^v)\big). \tag{2}$$

One the other side, the modality-specific networks are trained to confuse the discriminator with the opposite modal label by minimizing the adversarial loss

$$\min_{\theta_a,\theta_v} \mathcal{L}_{\text{adv}} = -\lambda_{\text{adv}}\mathbb{E}_{\mathbf{X},\mathbf{Y}\in\mathcal{D}} \sum_{i=1}^{N} \log\big(D(\hat{\mathbf{x}}_i^v)\big) + \log\big(1 - D(\hat{\mathbf{x}}_i^a)\big) + \log\big(D(\hat{\mathbf{y}}_i^v)\big) + \log\big(1 - D(\hat{\mathbf{y}}_i^a)\big) \tag{3}$$

where λ_{adv} is the weight coefficient. The discriminator guides the modal encoders to learn the same distribution representations through min–max adversarial learning.

2) Feature Fusion Attention: After we obtain the modality-invariant representations, we concatenate the audio and visual features for two inputs as $\mathbf{x}_f = [\hat{\mathbf{x}}_i^a, \hat{\mathbf{x}}_i^v]$ and $\mathbf{y}_f = [\hat{\mathbf{y}}_i^a, \hat{\mathbf{y}}_i^v]$, $[\cdot]$ denotes the concatenation operator. In particular, we design a fusion attention module to emphasize the efficient vector values. It consists of an MLP with the Sigmoid function and the output is the weight vector $\mathbf{w}$ with the same dimension of $\mathbf{x}_f$ and $\mathbf{y}_f$, which can be calculated by

$$\mathbf{w}_x = \sigma\big(\text{FCs}\big(\mathbf{x}_f\big)\big),\ \ \mathbf{w}_y = \sigma\big(\text{FCs}\big(\mathbf{y}_f\big)\big) \tag{4}$$
$$\mathbf{x} = \mathbf{x}_f\mathbf{w}_x,\ \ \mathbf{y} = \mathbf{y}_f\mathbf{w}_y \tag{5}$$

where $\sigma(\cdot)$ is the Sigmoid function and FCs$(\cdot)$ is the stacked two fully connected layers. The original concatenated feature $\mathbf{x}_f$ and $\mathbf{y}_f$ has the dimension of 2560. The first FC layer reduces the feature dimension to 1024, and then the last FC layer increases the dimension to the original 2560-D. After passing the Sigmoid activation, the weight vector could be obtained and we can obtain the adaptive feature fusion by using (5).

TABLE II

DATA STATISTICS FOR STUDYING THE AUDIO-VISUAL KINSHIP VERIFICATION IN THE WILD ON THE TALKIN-FAMILY DATASET. THE *# folds* MEANS THE NUMBER OF FOLD VALIDATIONS FOR EACH KIN RELATION. THE *# families and # subjects* REPRESENT HOW MANY FAMILIES AND INDIVIDUALS ARE INVOLVED WHEN STUDYING THE SPECIFIC KIN RELATION. THE *# kin pairs* MEANS THE NUMBER OF KIN PAIRS AT THE SUBJECT LEVEL. THE *# videos* IS THE TOTAL NUMBER OF VIDEOS USED, WHICH IS USUALLY FOUR TIMES THE NUMBER OF SUBJECTS, SINCE EACH SUBJECT HAS FOUR FACIAL VIDEOS. THE *# sample pairs* IS THE NUMBER OF FRAME-LEVEL SAMPLE PAIRS IN EACH KIN RELATION. APPLICABLE ALSO TO TABLE VIII

Generation	—Siblings—			—Parent-Child—				—Grandparent-Grandchild—			
Relations	BB	SS	BS	FS	FD	MS	MD	GFGS	GFGD	GMGS	GMGD
# folds	5	4	4	5	5	5	5	2	2	3	3
# families	24	33	31	86	62	134	125	9	10	14	12
# subjects	50	70	73	196	136	308	285	19	21	31	27
# kin pairs	200	336	320	848	576	1296	1264	80	88	136	120
# videos	200	280	292	784	544	1232	1140	76	84	124	108
# sample pairs	11570	17530	15124	45444	30379	71740	69926	4328	4724	6181	6250

$\mathbf{x}$ and $\mathbf{y}$ are the fused representations to obtain the kinship analysis. We denote attention parameters as θ_{att}.

D. Learning Kinship Awareness Embedding

To perceive the kinship traits, that is, similarity between kinship and difference between nonkinship, we adopt the contrastive learning to train the network in a supervised way. By integrating the kinship label l_i, the network objective can be expressed as

$$\min_{\theta_a,\theta_v,\theta_{\text{att}}} \mathcal{L}_{\text{kin}} = \frac{1}{2N}\sum_{i=1}^{N}\Big(l_i d^2 + (1-l_i)\max(M-d,0)^2\Big) \quad (6)$$

where threshold M is the margin, $d = \|\mathbf{x}-\mathbf{y}\|^2$.

The training procedure is summarized in Algorithm 1. During each training step, two multimodal encoders are first trained alternatively in an adversarial way together with discriminator without kin label evolved. Then, the entire network is jointly trained using the kin labels.

During the testing process, we collect the fused feature from the network. The cosine similarity $\text{sim}(\mathbf{x},\mathbf{y}) = ((\mathbf{x}\cdot\mathbf{y})/(\|\mathbf{x}\|\cdot\|\mathbf{y}\|))$ is calculated to represent the distance between two subjects. A threshold applied to *sim* determines whether two inputs have a kin relation as has been done in [21].

Algorithm 1: Training Procedure of Our UAAML

Input: Training set $\mathcal{D}$, initialize modality-specific encodes E_a, E_v, hyper-parameter λ_{adv}
Output: The parameters $\theta_a, \theta_v, \theta_{att}$

while not converged **do**
  for t-steps **do**
    update parameters θ_d of the discriminator by ascending their stochastic gradients:
    $\theta_d \leftarrow \theta_d - \eta \cdot \nabla_{\theta_d}\mathcal{L}_d$
    $\theta_a \leftarrow \theta_a - \eta \cdot \nabla_{\theta_a}\mathcal{L}_{adv}$
    $\theta_v \leftarrow \theta_v - \eta \cdot \nabla_{\theta_v}\mathcal{L}_{adv}$
  end for
  for d-steps **do**
    update parameters $\theta_a, \theta_v, \theta_{att}$ of the discriminator by ascending their stochastic gradients:
    $\theta_a \leftarrow \theta_a - \eta \cdot \nabla_{\theta_a}\mathcal{L}_{kin}$
    $\theta_v \leftarrow \theta_v - \eta \cdot \nabla_{\theta_v}\mathcal{L}_{kin}$
    $\theta_{att} \leftarrow \theta_{att} - \eta \cdot \nabla_{\theta_{att}}\mathcal{L}_{kin}$
  end for
end while
return $\theta_a, \theta_v, \theta_{att}$

V. EXPERIMENTS

A. Implementation Detail

1) Data Preparation: We first generate kin pairs with 11 relationship types described in Section III, where the sample pairs have different backgrounds and speak content. After we obtain the kinship pairs (positive pairs), we split them into maximum of five folds to conduct the K-fold validation [21]. Within each fold, we randomly generate the nonkinship pairs as negative samples, where nonkinship subjects are from different families and biologically unrelated. The negative samples have the same size as positive samples. Note that there is no family overlap between folds. The experimental data statistics distribution of audio-visual kinship verification in the wild is shown in Table II. The reason why it cannot be divided into five folds for relations, such as SS and BS, is that the negative samples suffer from insufficient families. We perform data preprocessing on all videos for visual and audio data as introduced in Section III. Since the video length varies from video to video and the neighbor video frames have a slight difference, we extract and align 60 facial frames and audio frames for each video. Due to the head variations and orientations, some frames are lost for a few subjects.

B. Compared Methods

To verify the effectiveness of our proposed method on the TALKIN-Family dataset and compare the performance between the unimodality and multimodalities, we perform baseline methods on vocal and FKV and four fusion methods.

1) Voice Features: We employ two methods: 1) GMM-UBM [59] and 2) I-vector [60], for audio analysis. We extract MFCCs with 12 cepstral coefficients from the audio samples. The UBM with 128 mixture components of GMM is trained with the training set. For the GMM-UBM [59] method, the kin pair model is created from UBM using the maximum a posteriori (MAP) estimation. The verification likelihood is the log-likelihood ratio between speaker models and registered speakers' GMM. In the I-vector [60] method, UBM is trained using expectation–maximization (EM) with MFCCs. The I-vector is obtained by MAP point estimation. Then, the dimension of the I-vector is reduced by linear discriminant analysis (LDA). We compute similarity between two speakers with the cosine similarity of I-vectors.

Besides, we also evaluate the pretrained deep models as feature encoders.

pyannote-S: The pyannote.audio [61], [62] is an End-to-End generic PyanNet that is trained on Voxceleb [24] and Voxceleb2 [23] datasets. The trained model takes the utterance and samples it with a sliding window to generate overlapping 512-D features. The pyannote-S means that we evaluate the performance using only the single vocal feature.

pyannote-A: For the utterance clip, we average all audio features for the sequence as its final feature representation.

VGG_M: The model architecture is based on VGG_M [24], and takes the audio spectrogram as input. The spectrogram is computed with the same method described in Section IV. VGG_M is trained on the Voxceleb dataset [24] with the task of speaker verification. The final audio feature has a length of 1024 dimensions.

ResNet-50: The model is trained on the Voxceleb2 dataset [23] and the audio embedding is collected from the FC layer with the length of 2048.

2) Facial Features: We consider four traditional facial descriptors: 1) BSIF [65]; 2) LPQ [64]; 3) LBP [63]; and 4) LBP-TOP [66], as has been done in [20]. We also implement the MNRML [3] metric-learning method that combines multiple feature descriptors, LBP, LPQ, and BSIF features, to learn the multiview data metric.

Furthermore, the deep CNN models pretrained on large-scale face datasets are also widely used in kinship verification to encode the facial image with output embedding.

SphereFace [67], [70] is a CNN model trained with the angular softmax (A-Softmax) to learn more discriminative features. The SphereFace is trained on the face dataset CASIA-WebFace [71]. Then, the deep features can be collected from the FC1 layer with 512 dimensions.

VGG-Face network [68] is trained on a large face dataset with 2.6 million images of over 2662 people. We feed the facial image into the network and collect features from layer fc7.

FaceNet-C [69], [72]: FaceNet is a deep CNN model trained with the Triple-let loss. FaceNet-C means the model trained on CASIA-WebFace [71]. The output feature is a 512-D embedding.

FaceNet-V [69], [72] means the FaceNet trained on the VGGFace2 [73] dataset.

InsightFace [56], [57]: Compared to SphereFace, InsightFace utilizes the AcFace loss that has fewer parameters yet with a better classification margin. The model is trained on the MS1MV2 dataset. The facial frames are fed into the pretrained model, and we can obtain the final 512-D feature embedding.

3) Fusion Methods: We perform both early fusion and two late fusion methods on audio-visual kinship verification.

Early Fusion: The multiview features are concatenated together as the fused feature for later similarity comparison.

Late Fusion (Mean): For the late fusion, the similarity scores are computed separately for each modality. Then, the mean fusion average scores were obtained from multimodalities as the final decision score.

Late Fusion (Max): Rather than calculating the averaged score, max fusion takes the maximum score as the final decision score.

Siamese fusion [20] introduced one FC layer to learn the fusion scheme. By adding the contrastive loss on the top of the network, the FC layer automatically learns the fusion weights for each element.

C. Experimental Settings

1) Implementation Details: We implement our network on the PyTorch library. Since the released pretrained InsightFace net and ResNet-50 (audio) are implemented based on MXNet and Matconvnet libraries, respectively. We first convert those models into PyTorch using opensource code from Github [74] and [75]. To initialize our network parameters, we use the ResNet-34 weights trained on MS1MV2 [56] for the visual network and the ResNet-50 weights trained on VoxCeleb2 [23] for the audio network. Parameters in other layers are initialized using random weights. For training the proposed method, the parameters of networks are optimized by the Adam optimizer with the learning rate of 1e-6, weight decay of 1e-4, and mini-batch size of 50. We train the entire network for 250 iterations. The program runs on two NVIDIA V100 GPUs (32 GB). The hyperparameter λ_{adv} determines the degree of multimodal discriminative information used during the model training process. In the case of using small λ_{adv}, no sufficient modality discrimination could be applied. We set λ_{adv} with 1 [76].

2) Evaluation Protocol: In our experiments, we compute the cosine similarity between two features. The threshold is used to classify whether two subjects have a kin relationship [21]. The verification accuracy and receiver operating characteristic (ROC) curves are used to evaluate the method's performance.

D. Experimental Results and Comparison

This section presents experimental results of kinship verification on the TALKIN-Family dataset from both single modality and multiple modalities.

1) Single-Modal Kinship Verification: Table III shows the kinship verification accuracy from single modality (based on one modality). For the voice-based kinship verification, the ResNet-50 has the best performance. The traditional methods I-vector and GMM-UBM have comparatively low performance. Notice that Grandparent–Grandchild results are not provided because the UBM is hard to converge due to limited data. The possible solution is to employ external data to train the UBM. Regarding the pyannote model, the performance can be slightly improved by averaging all vocal features within one utterance.

For kinship verification from faces, deep models outperform traditional descriptors by a large margin. Compared to traditional descriptors, the MRNML metric-learning method [3] has a better average accuracy, and the spatial–temporal descriptor LBP-TOP also outperforms the averaged frame-level features. Among deep learning models, InsightFace surpasses others with a large margin except for GFGS, that VGG-Face achieves

TABLE III
AVERAGE ACCURACIES (%) FOR K-FOLD KINSHIP VERIFICATION WITH VOICES, FACES, AND FUSION OF VOICES AND FACES UNDER THE WILD CONDITIONS IN THE TALKIN-FAMILY DATASET

Generation	—Siblings—			—Parent-Child—				—Grandparent-Grandchild—				Average
Relations	BB	SS	BS	FS	FD	MS	MD	GFGS	GFGD	GMGS	GMGD	
Vocal kinship verification												
I-vector [60]	63.11	65.99	60.99	63.07	63.23	61.22	62.17	-	-	-	-	-
GMM-UBM [59]	70.67	64.69	62.63	65.85	66.87	70.15	70.68	-	-	-	-	-
VGG-M [23]	68.83	64.21	59.33	62.23	56.18	56.94	57.99	57.13	59.38	73.01	65.83	61.91
pyannote-S [62]	71.61	61.40	65.08	61.22	55.63	59.21	58.67	64.85	57.50	62.54	63.61	61.94
pyannote-A [62]	72.11	65.70	63.95	63.58	58.44	57.93	59.27	67.13	57.92	71.60	58.61	63.30
ResNet-50 [24]	71.78	75.35	74.10	69.37	67.56	67.52	70.95	66.54	61.04	80.37	71.11	70.52
Facial kinship verification												
LBP [63]	69.67	63.71	56.52	60.23	59.13	59.67	60.86	55.44	62.29	63.65	58.33	60.86
LPQ [64]	64.72	60.50	61.79	57.38	60.55	60.84	63.58	60.44	55.21	64.20	67.50	61.52
BSIF [65]	69.44	66.97	65.34	59.71	63.01	62.98	61.95	62.13	54.17	69.20	67.50	63.85
LBP-TOP [66]	69.28	62.31	64.63	63.07	60.70	63.71	64.28	67.94	59.17	60.23	64.17	63.59
MNRML [3]	66.89	64.41	64.38	59.45	62.49	63.20	64.78	64.41	58.54	69.53	67.50	64.14
SphereFace [67]	70.11	62.90	59.98	58.75	57.93	59.61	56.91	59.41	60.21	62.53	73.33	61.97
VGG-Face [68]	75.11	63.71	63.08	62.36	59.29	63.67	63.16	68.60	61.67	63.33	71.67	65.06
FaceNet-C [69]	85.11	72.95	68.49	64.84	63.31	69.85	69.19	64.63	63.54	68.09	63.33	68.48
FaceNet-V [69]	86.61	74.82	65.08	68.03	68.15	67.63	68.04	67.35	65.00	66.56	63.33	69.15
InsightFace [56]	87.22	83.31	76.45	78.00	73.21	75.65	76.12	65.66	70.83	75.26	72.50	75.84
Audio-visual kinship verification using ResNet-50 and InsightFace as unimodal backbones												
UAAML (Proposed)	**90.02**	**86.95**	**78.30**	**80.74**	**77.17**	**76.97**	**76.65**	**70.48**	**78.67**	**83.18**	**73.47**	**79.33**

TABLE IV
COMPARISON OF DIFFERENT FUSION METHODS ON THE TALKIN-FAMILY DATASET FOR AUDIO-VISUAL KINSHIP VERIFICATION IN THE WILD WITH AVERAGE ACCURACIES (%) FOR K-FOLD VALIDATION. THE FIRST TWO ROWS ARE SINGLE-MODAL VERIFICATION PERFORMANCE WITH "A" SHORT FOR AUDIO AND "V" FOR VIDEO. APPLICABLE ALSO TO TABLE VII

Generation	—Siblings—			—Parent-Child—				—Grandparent-Grandchild—				Average
Relations	BB	SS	BS	FS	FD	MS	MD	GFGS	GFGD	GMGS	GMGD	
ResNet-50 (A) [24]	71.78	75.35	74.10	69.37	67.56	67.52	70.95	66.54	61.04	80.37	71.11	70.52
InsightFace (V) [56]	87.22	83.31	76.45	78.00	73.21	75.65	76.12	65.66	70.83	75.26	72.50	75.84
Late fusion (max)	81.56	81.52	**79.53**	77.05	71.90	73.35	75.81	68.02	68.33	**85.78**	71.94	75.89
Siamese fusion [20]	84.89	84.90	77.50	77.13	74.72	73.88	76.49	64.76	76.54	81.05	72.14	76.73
Early fusion	86.33	83.31	75.65	78.91	73.76	77.55	77.31	69.41	71.88	80.21	73.89	77.11
Late fusion (mean)	87.33	84.49	74.84	79.55	73.54	**77.97**	**77.46**	67.94	73.13	80.91	**78.33**	77.77
UAAML (Proposed)	**90.02**	**86.95**	78.30	**80.74**	**77.17**	76.97	76.65	**70.48**	**78.67**	83.18	73.47	**79.33**

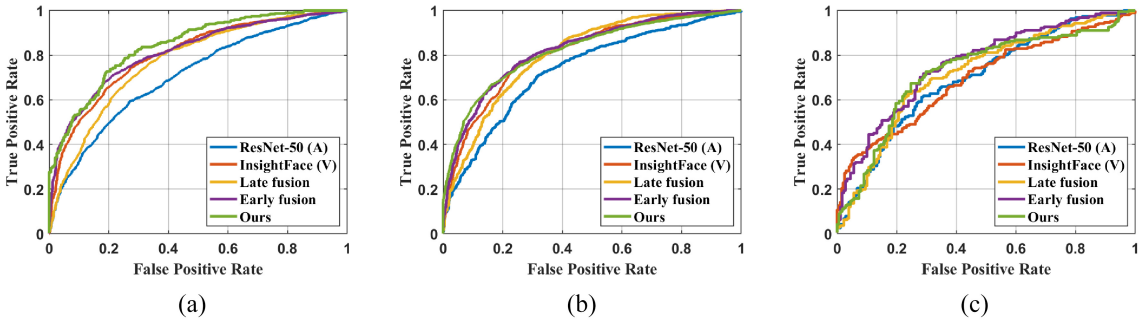

Fig. 3. ROC curves of different methods on TALKIN-Family with the wild condition obtained on (a) siblings, (b) parent–child, and (c) grandparent–grandchild kin relations.

the best performance. The better models boost the kinship verification performance due to the accurate feature representations. Therefore, we apply ResNet-50 (voice) and InsightFace (face) as the backbone networks for the fusion.

2) Multimodalities: As presented at the end of Table III, the proposed UAAML method shows an improvement over the single modalities for all 11 kin relations and the average level. Table IV also compares the results of several baseline fusion methods. Fig. 3 visualizes different methods' corresponding ROC curves. It can be seen that by fusing the audio and visual features, the performance could be improved, demonstrating that the vocal and facial features complement each other. In addition, the proposed fusion method improves the single-modality verification accuracies and the baseline fusion methods to a certain extent. Average accuracy improves by about 3.5% and 2.0% from the single modality and baseline fusion methods. Although baseline fusion methods cannot beat the UAAML method at the average level, score fusion methods show slightly higher accuracy in relations such as BS, NS, and GMGS. This is a motivation for future work that further explores multifusion strategies for audio-visual kinship verification.

3) Ablation Study: To analyze the effect of different components of UAAML, we ablate the proposed method and evaluate the effectiveness of each.

a) Fusing different features: To study fusing various single-modality features, we include the two single-modal features for both face and voice modality into evaluation. The VGG-M and ResNet-50 models are used for vocal features, and the FaceNet-V and InsightFace models are applied

TABLE V
COMPARISON BY FUSING DIFFERENT SINGLE-MODAL FEATURES. (*A1* IS THE VOCAL FEATURE COLLECTED FROM RESNET-50 TRAINED ON VOXCELEB2, *A2* IS THE VOCAL FEATURE OBTAINED FROM VGG_M TRAINED ON VOXCELEB; *V1* IS THE FACIAL FEATURE EXTRACTED FROM INSIGHTFACE, AND *V2* IS THE FACIAL FEATURE COLLECTED FROM FACENET-V)

Vocal features		Facial features		Average (%)
A1	A2	V1	V2	
✓				70.52
	✓			61.91
		✓		75.84
			✓	69.15
✓	✓			67.84
		✓	✓	72.63
✓			✓	70.92
	✓	✓		74.76
✓	✓	✓		74.97
✓		✓	✓	74.92
✓	✓	✓	✓	74.30
✓		✓		**77.11**

for facial features. To simplify the process, we implement feature fusion to evaluate the effectiveness of the fusion. The L_2 normalization is computed before fusion to reduce the discrepancy within different features. Table V shows the averaged verification accuracy when combining various multiple features. The experimental results show that the InsightFace (Face) and ResNet-50 (Voice) feature fusion achieves the best performance. However, when combining VGG-M (voice) features or FaceNet-V features with comparatively low performance, the system can be easily affected by poor features. Therefore, the InsightFace and ResNet-50 encoders are used as our backbone networks.

b) Roles of different losses and components: We further evaluate the effectiveness of adversarial learning, contrastive learning loss, and attention layer: 1) w/o. att + L_{kin} denotes the network discards the adversarial learning and the attention layer, and it is trained with the contrastive learning loss; 2) w/o. att + L_{adv} denotes the adversarial network without the attention layer, which is trained with the self-supervised learning strategy [77] without kin labels. It learns the consistency between modalities to embed the semantic multimodal features; and 3) w/ att + L_{kin} denotes the network discards the adversarial learning module but keeps the attention layer, which is trained with the kinship loss. Table VI reports the verification accuracy. Experimental results demonstrate the necessity of the adversarial module, attention layer, and kinship loss. The proposed UAAML further improves the performance compared with the three variants. Those results also convey that adversarial and attention modules are the key components for audio-visual kinship verification.

c) Normalization layer: We perform the model training with the same efforts without the normalization layer. As shown in Fig. 4, the performance drops significantly, showing that the normalization layer is crucial to make the training process stable and improve the performance.

E. Evaluation on the TALKIN Dataset

In this section, we further evaluate the effectiveness of the proposed UAAML method on the TALKIN dataset for audio-visual kinship verification. The TALKIN dataset has four parent–child kin relations, that is, FS, FD, MS, and MD. For each kin relation, there are 100 pairs of kin facial videos, and 100 pairs of nonkin videos. The five-fold validation is performed. Similarly to previous experimental settings, we apply the InsightFace with ResNet-34 architecture [56], [57] (face) and ResNet-50 [23], [58] (voice) as the single-modality backbone networks. Table VII presents the performance of single-modality methods and different fusion methods, and Fig. 5 shows the corresponding ROC curves. The experimental results demonstrate that the proposed UAAML method obtains the highest level of accuracy compared to both single-modality and baseline fusion methods. The baseline score fusion method (max) shows a 1.0% higher accuracy in the FS relation compared with UAAML. Considering that the videos in TALKIN contain additive background noise, the performance of the audio modality is relatively worse and, thus, brings limited fusion improvement. Therefore, for audio-visual kinship verification, especially when it comes to real-world problems, more robust voice models are needed [78].

TABLE VI
LOSS AND MODULE ANALYSIS OF THE UAAML METHOD ON THE TALKIN-FAMILY DATASET. THE ATT IS THE ABBREVIATION OF FEATURE ATTENTION

Models	Average (%)
Basic-voice	70.52
Basic-face	75.84
w/o. att + L_{kin}	76.73
w/o. att + L_{adv}	78.55
w/ att + L_{kin}	78.90
w/ att + L_{kin} + L_{adv}	**79.33**

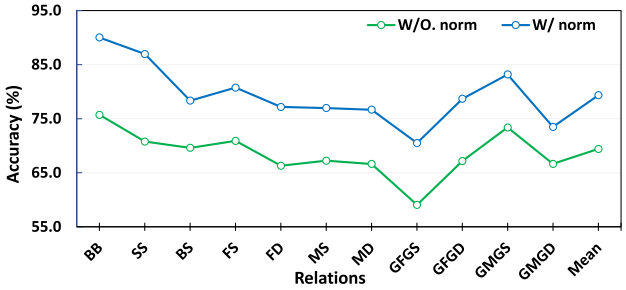

Fig. 4. Comparison of the effect when we take the same efforts training the network with normalization and without normalization.

TABLE VII
COMPARISON OF DIFFERENT FUSION METHODS ON THE TALKIN DATASET WITH AVERAGE ACCURACIES (%) FOR FIVE-FOLD VALIDATION

Generation	—Parent-Child—				Average
Relations	FS	FD	MS	MD	
ResNet-50 (A) [24]	69.50	59.50	63.50	65.50	64.50
InsightFace (V) [56]	83.00	74.50	75.50	80.50	78.38
Siamese fusion [20]	84.00	72.50	76.50	80.50	78.38
Late fusion (mean)	83.00	74.50	76.00	82.00	78.88
Early fusion	83.00	74.50	76.00	81.50	78.75
Late fusion (max)	**84.50**	75.00	77.00	80.00	79.13
UAAML(Proposed)	83.50	**75.50**	**77.50**	**82.00**	**79.63**

F. Influence Factors

The audio-visual kinship verification is affected by many factors. From the perspective of biological attributes, this

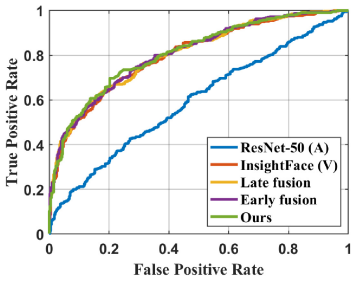

Fig. 5. ROC curves of different methods on the TALKIN dataset obtained on parent–child kin relations.

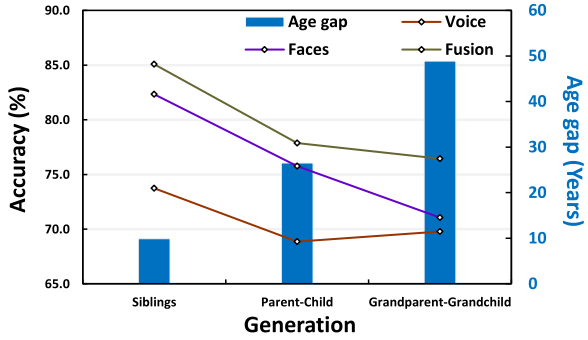

Fig. 6. Line charts illustrate the verification accuracy on different modalities. The bar chart shows the age gap between the kin subjects.

includes the depth of the genealogical tree, age, and gender. From the data acquisition conditions, the factors include the recording background and video speech content. We analyze how those factors influence the performance by providing the corresponding experimental results.

1) Genealogical Tree: Fig. 6 shows the averaged verification accuracy for three generations of kinship with different inputs. It can be seen that the deeper the genealogical tree, the performance on faces drops significantly. One reason for this is the age difference between kinship, as distributed in Fig. 6. The siblings of the same generation have the smallest age difference of about ten years on average, of which parent–child has about a 26-year age difference. However, second-generation subjects have an average age difference of about 50 years. As people start aging, the appearance of their faces varies in structure and texture. These differences affect the inner similarity of the kin image pairs, consequently reducing the verification performance, whereas acoustic features compensate for facial aging issues to some extent, especially for the Grandparent–Grandchild relationship.

2) Gender Factor: The experimental setting of relation-specific evaluation provides us with the possibility of analyzing the influence brought by gender. From Table IV, we could observe that the influence of gender is significant for siblings, where the opposite gender (BS) has a comparatively lower accuracy than the cases with the same gender (BB, SS). Regarding the parent–child and grandparent–grandchild relations, the influence of gender is more limited, and its impact is lower than the influence caused by the texture difference brought by the age gap. On the other hand, on the TALKIN dataset, the influence of opposite genders can be found in the

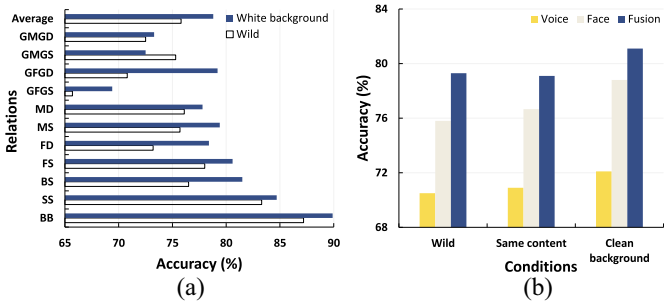

Fig. 7. Performance of kinship verification on TALKIN-Family under different conditions. (a) Shows the performance comparison on the visual kinship verification under white and nonwhite backgrounds. (b) Compares the single modal and multimodal performance with different data recording settings.

parent–child relations (Table VII), as some kinship videos are recorded at the similar age (e.g., FS pairs: [25, 26], [43, 44], [45, 46], [71, 72], etc.), rather than at the same time (e.g., TALKIN-Family).

3) Recording Conditions: The data collection conditions potentially influence the system performance, such as speech text in speaker verification [51], and the same photo issue in kinship verification [50] by providing latent clues. To control one variable factor for one time, we generate the kinship pairs that: 1) speak the fixed text but with different backgrounds (text-dependent) and 2) are recorded under the white background but with different speaking content (white background). The data statistics on the two scenarios are listed in Table VIII. Fig. 7 shows the experimental results on text-dependent and white background conditions with different inputs. The background influence could be clearly seen from Fig. 7(a) that the white background performance has higher accuracy. Two reasons explain the phenomenon: 1) the noise effect is eased under the white background and 2) the white background videos within one family are possibly recorded at the same place, with similar illumination, which could cause data bias [50]. This also explains why we asked the participants to take videos under two backgrounds, one of which is white, to easily distinguish the same or different backgrounds. As illustrated in Fig. 7(b), the fixed text setting achieves comparable performance to the free-speaking setting due to the equal similarity within kin and nonkin pairs. Overall, the audio-visual fusion improves performance under all conditions, while under two semicontrolled environments, the improvement of fusion is comparably limited.

G. Human Performance

We test the human performance on kinship verification by using a subset of TALKIN-Family. Twelve volunteers from China participated in the experiments. Before the test, they had never seen or known any information about the dataset subjects. They were asked to answer whether the given clips have a kin relation. In general, we set up three types of tasks, namely, kinship verification from: 1) *facial videos without voice*; 2) *voice*; and 3) *facial videos with voice*. For each task, we select two kin pairs and two nonkin pairs from each of 11 kin relations, resulting in 22 positive pairs (kinship) and 22 negative pairs (nonkinship) in total. To avoid

TABLE VIII
DATA STATISTICS FOR THE AUDIO-VISUAL KINSHIP VERIFICATION UNDER CONDITIONS OF FIXED SPEECH AND CLEAN BACKGROUND ON THE TALKIN-FAMILY DATASET

Generation	—Siblings—			—Parent-Child—				—Grandparent-Grandchild—			
Relations	BB	SS	BS	FS	FD	MS	MD	GFGS	GFGD	GMGS	GMGD
With the same speaking content											
# kin pairs	100	168	160	424	288	648	632	40	44	68	60
# sample pairs	5821	8809	7408	22988	14809	36321	34776	2240	2331	3216	3100
Under the clean background											
# kin pairs	100	168	160	424	288	648	632	40	44	68	60
# sample pairs	5826	8622	7516	22806	14980	36587	34817	2162	2350	3040	3106

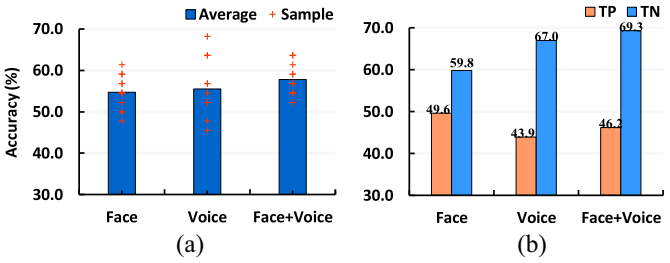

Fig. 8. Human performance on a subset of TALKIN-Family from the face, voice, and face&voice, respectively: (a) overall verification performance with different modalities and (b) TP and TN distributions of human performance under different settings.

the recall of previously seen information, we designed the set such as there is no subject overlap between positive and negative pairs or among the three subtasks. Fig. 8 illustrates the human performance results, in which Fig. 8(a) shows the overall accuracy and distribution of the subject performance. We compare the true positive (TP) and true negative (TN) accuracy in Fig. 8(b). Generally, an important finding is that humans tend to have a better ability to verify kinship from voice than from face, while when given synchronous facial videos and voice, humans can make a better judgment. Fig. 8(a) indicates that face and voice information enables human observers to make a more stable assessment and higher accuracy. Fig. 8(b) shows that the humans have higher accuracy in verifying the negative samples, and multimodal information helps humans to recognize nonkinship, thus improving the overall accuracy. It is worth noting that it takes about an hour for one observer to complete the entire test, while machine-learning methods spend much less time in the inference process. We conclude that machine-learning methods can outperform human ability both efficiently and effectively.

VI. CONCLUSION AND FUTURE WORK

Audio-visual kinship verification is a new and potential research topic. In this article, we systematically investigate the problem of audio-visual kinship verification. We establish the most comprehensive audio-visual kinship dataset, called TALKIN-Family. Moreover, the baseline experiments of single-modal kinship verification are performed, of which the vocal kinship verification is evaluated for the first time. Based on the single-modal methods, we provide a deep learning framework, called UAAML, to jointly learn the modal-invariant and adaptive fused features for kinship verification with contrastive loss. The extensive experimental results demonstrate the effectiveness of audio-visual fusion compared to unimodal methods. Our proposed fusion method could outperform to the baseline methods. The human performance shows that by providing both the faces and voices, people could have higher kinship verification accuracy than using faces or voices only.

We expect this work sets a milestone for audio-visual kinship verification. To stimulate future study, in this section, we investigate the limitations of our datasets and the proposed approach and discuss future directions. Finally, we point out how TALKIN-Family can be applied in research beyond kinship verification.

A. Limitations and Future Work

1) TALKIN-Family Dataset: The offline data collection has drawbacks, such as the difficulty of increasing the data volumes, the cost of human effort to collect the data, and homogeneous ethnicity distributions. Given this, the future work is considered to speed up the data collection procedure by applying crowdsourcing, at the same time, saving manual labor and increasing the data diversity. Since the TALKIN-Family only has people from China, when conducting the validation on other ethnicities, the ethnicity adaptation and how to mitigate the demographic bias [79] can be a future research direction.

2) UAAML: The main limitation of the proposed UAAML is that the model training demands high computational resources. However, during the inference time, the proposed method is comparable to simpler methods such as the naive fusion. In our experimental results, the late fusion shows better performance in some kin relations. We argue that the reason lies in the different scores that are better classified by the late fusion. This inspires us to explore hybrid fusion methods in the future to combine the advantages of both. More effective and efficient fusion methods are demanded for audio-visual kinship verification, such as multimodal regularization [80], and multimodal joint representation [81] learning the complementary semantics.

B. Research Opportunities With the TALKIN-Family Dataset

This work focuses on studying the audio-visual kinship verification based on the TALKIN-Family dataset. Beyond it, the proposed dataset could also be used in studying kinship with a wide range. The TALKIN-Family database contains family information, subject labels, environment context, etc. Those data attributes allow researchers to explore kinship verification with intensive analysis for example, at the family level, on the

effects of age and gender, and with background context and speaking content. Based on audio-visual kinship verification, the study could also be extended to other kinship recognition problems, such as trisubject kinship verification [30], family recognition, family retrieval [21], child face/voice generation. Furthermore, the robustness of multimodal kinship recognition is also an open issue, such as against adversarial attack [82], spoof attack [83], [84], poor conditions (e.g., modality missing and cross-modal feature learning). Data bias, fairness [79], and privacy-aware studies [85] are also worthy of further attention with the growing concern of data privacy protection. TALKIN-Family can also be helpful in audio-visual studies, such as talking face generation [86] and face-voice matching [87], and human perception studies on kin faces and voices.

In conclusion, we expect that TALKIN-Family could motivate researchers from different fields to advance the audio-visual kinship studies, techniques, and applications and enable further development.

ACKNOWLEDGMENT

The authors would like to give special thanks to all volunteers for their contribution to the TALKIN-Family dataset. The data collection help from An Huang is also acknowledged. They would also like to express their sincere appreciation to the Associate Editor and the anonymous reviewers for their comments and suggestions. The CSC-IT Center for Science, Finland, is acknowledged for providing computational resources.

REFERENCES

[1] X. Wu et al., "Facial kinship verification: A comprehensive review and outlook," *Int. J. Comput. Vis.*, vol. 130, pp. 1494–1525, Apr. 2022.
[2] R. Fang, K. D. Tang, N. Snavely, and T. Chen, "Towards computational models of kinship verification," in *Proc. ICIP*, 2010, pp. 1577–1580.
[3] 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.
[4] L. Zhang, Q. Duan, D. Zhang, W. Jia, and X. Wang, "AdvKin: Adversarial convolutional network for kinship verification," *IEEE Trans. Cybern.*, vol. 51, no. 12, pp. 5883–5896, Dec. 2021.
[5] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, Mar. 2021.
[6] H. Dibeklioglu, "Visual transformation aided contrastive learning for video-based kinship verification," in *Proc. ICCV*, 2017, pp. 2459–2468.
[7] S. Huang, J. Lin, L. Huangfu, Y. Xing, J. Hu, and D. D. Zeng, "Adaptively weighted k-tuple metric network for kinship verification," *IEEE Trans. Cybern.*, early access, Apr. 19, 2022, doi: 10.1109/TCYB.2022.3163707.
[8] F. Hansen, L. M. DeBruine, I. J. Holzleitner, A. J. Lee, K. J. O'Shea, and V. Fasolt, "Kin recognition and perceived facial similarity," *J. Vis.*, vol. 20, no. 6, p. 18, 2020.
[9] R. T. Sataloff, "Genetics of the voice," *J. Voice*, vol. 9, no. 1, pp. 16–19, 1993.
[10] M. Weirich and L. Lancia, "Perceived auditory similarity and its acoustic correlates in twins and unrelated speakers," in *Proc. ICPhS*, 2011, pp. 2118–2121.
[11] S. P. Whiteside and E. Rixon, "Speech tempo and fundamental frequency patterns: A case study of male monozygotic twins and an age-and sex-matched sibling," *Logopedics Phoniatrics Vocol.*, vol. 38, no. 4, pp. 173–181, 2013.
[12] F. Debruyne, W. Decoster, A. Van Gijsel, and J. Vercammen, "Speaking fundamental frequency in monozygotic and dizygotic twins," *J. Voice*, vol. 16, no. 4, pp. 466–471, 2002.
[13] F. Nolan, K. McDougall, and T. Hudson, "Some acoustic correlates of perceived (Dis) similarity between same-accent voices," in *Proc. ICPhS*, 2011, pp. 1506–1509.
[14] W. G. Van, J. Vercammen, and F. Debruyne, "Voice similarity in identical twins," *Acta Oto-Rhino-Laryngologica Belgica*, vol. 55, no. 1, pp. 49–55, 2001.
[15] S. M. Taylor, "Acoustic correlates of aging and familial relationship," Ph.D. dissertation, Dept. Commun. Disorders, Brigham Young Univ., Provo, UT, USA, 2018.
[16] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 60–75, Jan.–Mar. 2019.
[17] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8717–8727, Dec. 2022.
[18] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy, "Learning to track and identify players from broadcast sports videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1704–1716, Jul. 2013.
[19] P. S. Aleksic and A. K. Katsaggelos, "Audio-visual biometrics," *Proc. IEEE*, vol. 94, no. 11, pp. 2025–2044, Nov. 2006.
[20] X. Wu, E. Granger, T. H. Kinnunen, X. Feng, and A. Hadid, "Audio-visual kinship verification in the wild," in *Proc. ICB*, 2019, pp. 1–8.
[21] J. P. Robinson, M. Shao, Y. Wu, H. Liu, T. Gillis, and Y. Fu, "Visual kinship recognition of families in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2624–2637, Nov. 2018.
[22] R. Vergin, D. O'Shaughnessy, and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 525–532, Sep. 1999.
[23] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. INTERSPEECH*, 2018, pp. 1086–1090.
[24] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. INTERSPEECH*, 2017, pp. 2616–2620.
[25] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[26] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. CVPR*, 2016, pp. 3640–3649.
[27] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," 2015, *arXiv:1506.04579*.
[28] M. Shao, S. Xia, and Y. Fu, "Genealogical face recognition based on UB KinFace database," in *Proc. CVPR WORKSHOPS*, 2011, pp. 60–65.
[29] R. Fang, A. C. Gallagher, T. Chen, and A. Loui, "Kinship classification by modeling facial feature heredity," in *Proc. ICIP*, 2013, pp. 2983–2987.
[30] X. Qin, X. Tan, and S. Chen, "Tri-subject kinship verification: Understanding the core of a family," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1855–1867, Oct. 2015.
[31] N. Kohli, M. Vatsa, R. Singh, A. Noore, and A. Majumdar, "Hierarchical representation learning for kinship verification," *IEEE Trans. Image Process.*, vol. 26, pp. 289–302, 2017.
[32] H. Dibeklioglu, A. A. Salah, and T. Gevers, "Like father, like son: Facial expression dynamics for kinship verification," in *Proc. ICCV*, 2013, pp. 1497–1504.
[33] H. Yan and J. Hu, "Video-based kinship verification using distance metric learning," *Pattern Recognit.*, vol. 75, pp. 15–24, Mar. 2018.
[34] N. Kohli, D. Yadav, M. Vatsa, R. Singh, and A. Noore, "Supervised mixed norm autoencoder for kinship verification in unconstrained videos," *IEEE Trans. Image Process.*, vol. 28, pp. 1329–1341, 2019.
[35] J. P. Robinson, Z. Khan, Y. Yin, M. Shao, and Y. Fu, "Families in wild multimedia: A multimodal database for recognizing kinship," *IEEE Trans. Multimedia*, vol. 24, pp. 3582–3594, Aug. 2021.
[36] X. Wu, E. Boutellaa, M. B. López, X. Feng, and A. Hadid, "On the usefulness of color for kinship verification from face images," in *Proc. WIFS*, 2016, pp. 1–6.
[37] F. Xiong, Y. Xiao, Z. Cao, Y. Wang, J. T. Zhou, and J. Wu, "ECML: An ensemble cascade metric-learning mechanism toward face verification," *IEEE Trans. Cybern.*, vol. 52, no. 3, pp. 1736–1749, Mar. 2022.
[38] K. Zhang, Y. Huang, C. Song, H. Wu, and L. Wang, "Kinship verification with deep convolutional neural networks," in *Proc. BMVC*, 2015, pp. 1–12.
[39] L. Li, X. Feng, X. Wu, Z. Xia, and A. Hadid, "Kinship verification from faces via similarity metric based convolutional neural network," in *Proc. ICIAR*, 2016, pp. 539–548.
[40] E. Dahan and Y. Keller, "A unified approach to kinship verification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2851–2857, Aug. 2021.
[41] W. Li, Y. Zhang, K. Lv, J. Lu, J. Feng, and J. Zhou, "Graph-based kinship reasoning network," in *Proc. ICME*, 2020, pp. 1–6.

[42] W. Li, J. Lu, A. Wuerkaixi, J. Feng, and J. Zhou, "Reasoning graph networks for kinship verification: From star-shaped to hierarchical," *IEEE Trans. Image Process.*, vol. 30, pp. 4947–4961, 2021.

[43] W. Li, S. Wang, J. Lu, J. Feng, and J. Zhou, "Meta-mining discriminative samples for kinship verification," in *Proc. CVPR*, 2021, pp. 16135–16144.

[44] C. Song and H. Yan, "KINMIX: A data augmentation approach for kinship verification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2020, pp. 1–6.

[45] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[46] X. Huang, Y. Peng, and M. Yuan, "MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1047–1059, Mar. 2020.

[47] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," in *Proc. AAAI*, vol. 34, 2020, pp. 164–172.

[48] R. Zhou and Y.-D. Shen, "End-to-end adversarial-attention network for multi-modal clustering," in *Proc. CVPR*, 2020, pp. 14619–14628.

[49] E. Tartaglione, C. A. Barbano, and M. Grangetto, "EnD: Entangling and disentangling deep representations for bias correction," in *Proc. CVPR*, 2021, pp. 13508–13517.

[50] M. B. López, E. Boutellaa, and A. Hadid, "Comments on the 'kinship face in the wild' data sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2342–2344, Nov. 2016.

[51] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. ICASSP*, 2016, pp. 5115–5119.

[52] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, 2017, pp. 999–1003.

[53] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

[54] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. CVPR*, 2014, pp. 1867–1874.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[56] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019, pp. 4690–4699.

[57] J. Guo and J. Deng. "InsightFace." 2021. [Online]. Available: https://github.com/deepinsight/insightface

[58] J. S. Chung, A. Nagrani, and A. Zisserman. "VoxCeleb2." 2017. [Online]. Available: https://github.com/a-nagrani/VGGVox

[59] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, nos. 1–3, pp. 19–41, 2000.

[60] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[61] H. Bredin et al. "Pyannote.audio." 2017. [Online]. Available: https://github.com/clcarwin/sphereface_pytorch

[62] H. Bredin et al., "Pyannote.audio: Neural building blocks for speaker diarization," in *Proc. ICASSP*, 2020, pp. 7124–7128.

[63] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[64] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[65] J. Kannala and E. Rahtu, "BSIF: Binarized statistical image features," in *Proc. ICPR*, 2012, pp. 1363–1366.

[66] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[67] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. CVPR*, 2017, pp. 6738–6746.

[68] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015, pp. 1–12.

[69] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015, pp. 815–823.

[70] Carwin. "SphereFace." 2020. [Online]. Available: https://github.com/pyannote/pyannote-audio-hub

[71] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*.

[72] T. Esler. "FaceNet." 2021. [Online]. Available: https://github.com/timesler/facenet-pytorch

[73] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. FG*, 2018, pp. 67–74.

[74] E. Nizhibitsky. "Pytorch-insightface." 2019. [Online]. Available: https://github.com/nizhib/pytorch-insightface

[75] S. Albanie. "Pytorch-mcn." 2018. [Online]. Available: https://github.com/albanie/pytorch-mcn

[76] P. Hu, D. Peng, X. Wang, and Y. Xiang, "Multimodal adversarial network for cross-modal retrieval," *Knowl. Based Syst.*, vol. 180, pp. 38–50, Sep. 2019.

[77] G.-N. Dong, C.-M. Pun, and Z. Zhang, "Deep collaborative multimodal learning for unsupervised kinship estimation," *IEEE Trans. Image Process.*, vol. 16, pp. 4197–4210, 2021.

[78] S. Kataria, P. S. Nidadavolu, J. Villalba, N. Chen, P. Garcia-Perera, and N. Dehak, "Feature enhancement with deep feature losses for speaker verification," in *Proc. ICASSP*, 2020, pp. 7584–7588.

[79] M. Georgopoulos, J. Oldfield, M. A. Nicolaou, Y. Panagakis, and M. Pantic, "Mitigating demographic bias in facial datasets with style-based multi-attribute transfer," *Int. J. Comput. Vis.*, vol. 129, no. 7, pp. 2288–2307, 2021.

[80] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[81] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[82] Y. Tian and C. Xu, "Can audio-visual integration strengthen robustness under multimodal attacks?" in *Proc. CVPR*, 2021, pp. 5601–5611.

[83] J. Y. Sim, H. W. Noh, W. Goo, N. Kim, S.-H. Chae, and C.-G. Ahn, "Identity recognition based on bioacoustics of human body," *IEEE Trans. Cybern.*, vol. 51, no. 5, pp. 2761–2772, May 2021.

[84] G. Krishnamurthy, N. Majumder, S. Poria, and E. Cambria, "A deep learning approach for multimodal deception detection," 2018, *arXiv:1803.00344*.

[85] C. Kumar, R. Ryan, and M. Shao, "Adversary for social good: Protecting familial privacy through joint adversarial attacks," in *Proc. AAAI*, vol. 34, 2020, pp. 11304–11311.

[86] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proc. AAAI*, vol. 33, 2019, pp. 9299–9306.

[87] S. Horiguchi, N. Kanda, and K. Nagamatsu, "Face-voice matching using cross-modal embeddings," in *Proc. ACM MM*, 2018, pp. 1011–1019.