IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 June 2024, accepted 3 July 2024, date of publication 11 July 2024, date of current version 22 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3426602

RESEARCH ARTICLE

One Model to Rule Them All: A Universal Transformer for Biometric Matching

MADINA ABDRAKHMANOVA, (Member, IEEE), ASSEL YERMEKOVA, YULIYA BARKO, VLADISLAV RYSPAYEV, MEDET JUMADILDAYEV, AND HUSEYIN ATAKAN VAROL, (Senior Member, IEEE)

Institute of Smart Systems and Artificial Intelligence, Nazarbayev University, Astana 010000, Kazakhstan

Corresponding author: Huseyin Atakan Varol (ahvarol@nu.edu.kz)

ABSTRACT This study introduces the first single branch network designed to tackle a spectrum of biometric matching scenarios, including unimodal, multimodal, cross-modal, and missing modality situations. Our method adapts the prototypical network loss to concurrently train on audio, visual, and thermal data within a unified multimodal framework. By converting all three data types into image format, we employ the Vision Transformer (ViT) architecture with shared model parameters, enabling the encoder to transform input modalities into a unified vector space. The multimodal prototypical network loss function ensures that vector representations of the same speaker are proximate regardless of their original modalities. Evaluation on SpeakingFaces and VoxCeleb datasets encompasses a wide range of scenarios, demonstrating the effectiveness of our approach. The trimodal model achieves an Equal Error Rate (EER) of 0.27% on the SpeakingFaces test split, surpassing all previously reported results. Moreover, with a single training, it exhibits comparable performance with unimodal and bimodal counterparts, including unimodal audio, visual, and thermal, as well as audio-visual, audio-thermal, and visual-thermal configurations. In cross-modal evaluation on the VoxCeleb1 test set (audio versus visual), our approach yields an EER of 24.1%, again outperforming state-of-the-art models. This underscores the effectiveness of our unified model in addressing diverse scenarios for biometric verification.

INDEX TERMS Biometric matching, cross-modal matching, face verification, face-audio association, metric learning, multimodal verification, speaker verification, transformer.

I. INTRODUCTION

Biometric matching is the process of verifying a person's identity based on the person's unique biological or behavioral characteristics (see Fig. 1). Specifically, person verification confirms an individual's identity by comparing and matching their biometric data with the data of that person previously stored in a system. Whereas, in person identification, an individual's biometrics are compared with the data of various other individuals to find a match, essentially identifying the individual in a larger group. Common biometric traits or modalities include face, voice, fingerprint, iris scans, and others [1].

The advancements in deep learning have significantly improved the performance of biometric systems for different modalities. Deep learning has shown impressive outcomes in facial recognition due to its ability to extract features from a large amount of data [2]. Regarding voice recognition, deep learning models based on recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have been effectively utilized as reported by Bai and Zhang [3]. CNNs (AlexNet, GoogLeNet, and ResNet) applied in fingerprint recognition for large databases achieved superior results compared to traditional methods [4]. Liu et al. [5] presented a novel condensed 2-channel CNN for efficient and accurate iris identification. Using a multi-branch model and fast fine-tuning improved performance while reducing computational complexity. Recently, leveraging deep learning, even brain signals were utilized for biometrics.

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

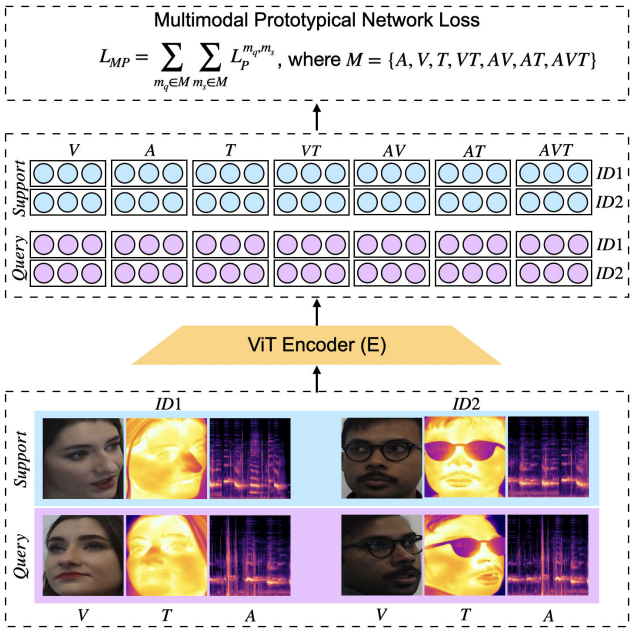

FIGURE 1. The training pipeline of our unified biometric matching system using the multimodal prototypical network loss on the audio-visual-thermal SpeakingFaces dataset. Each speaker is represented by support and query samples across three modalities: log-mel spectrogram (A), thermal facial image (T), and visual facial image (V). A shared Vision Transformer (ViT) encoder extracts embeddings from each modality, visualized as light blue (support) and purple (query) circles. The extracted unimodal vectors are then fused to produce multimodal combinations (VT, AV, AT, AVT). The multimodal prototypical network loss clusters embeddings of the same speaker together in a shared vector space, minimizing intra-class distances and maximizing inter-class distances to enable universal biometric matching (unimodal, cross-modal, multimodal).

For instance, a study explored the effectiveness of deep learning techniques, namely CNNs and RNNs, in extracting distinctive features from electroencephalogram signals, demonstrating high-level accuracy for brain-based biometric recognition under challenging conditions across extended time periods [6].

The categorization of biometric matching depends on the number of available modalities, leading to distinctions such as unimodal, multimodal, and cross-modal. Unimodal matching involves using a single biometric trait for identity verification. Exceptional results have been achieved with face [7], [8], [9] and voice biometrics [3], [10]. Although unimodal systems have the advantage of simplicity and ease of implementation, they may be vulnerable to challenging conditions. For example, a person might be in a low-light environment or partially occluded, resulting in the visual stream being unreliable for verification. In other cases, the audio track might be corrupted due to background noise or microphone malfunction.

Multimodal biometric matching takes advantage of multiple types of data simultaneously, while compensating or addressing the limitations that individual modalities may have on their own [11]. Such approach has shown to be superior not only in accuracy, but also in reliability [12], [13], [14]. A notable drawback of multimodal systems is their dependence on the presence of all modalities. They tend to encounter difficulties and performance degradation when faced with missing or corrupted modalities.

Cross-modal biometric matching extends the concept of multimodal matching by comparing data from different biometric modalities [15]. It can be especially useful in scenarios with incomplete multimodal data due to sensor malfunctions, data corruption, and enviromental noise. Recently, cross-modal processing has been applied in various combinations, such as audio-visual [14], [16], [17], [18], [19], [20], [21], [22], [23], thermal-visual [24], [25], [26], and audio-text [27]. The common strategy employed in these studies involves mapping inputs from diverse modalities into a shared space to facilitate cross-modal retrieval. For instance, in [16], a system, with separate subnetworks for each modality, is employed to extract low-dimensional embedding vectors from speech and facial data. These subnetworks are trained using a contrastive loss function to map matching face and voice embeddings to the same space. Sari et al. [14] introduced a multi-view system designed to produce high-level representations for audio and video modalities within a shared space that spans both modalities. This was accomplished by employing a shared classifier for the outputs of the audio and video encoders, ensuring that when optimized together, the encoder outputs were projected into a common space.

All of the multimodal and cross-modal approaches above involve using separate encoders to extract feature embeddings for each modality, that can be fused at different stages to achieve joint representation. However, it is worth noting that embeddings derived from modality-specific networks often exhibit significant semantic similarities. For instance, attributes like gender, ethnicity, and age of speakers are reflected in both their audio and visual signatures [16]. In [17], a single stream network (SSNet) was trained with a new loss function to map audio and visual embeddings into a shared latent space, while maintaining neighborhood constraints within and across the two modalities. Saeed et al. [21] introduced a single-branch network (SBNet), designed to acquire a discriminative representation for both unimodal and multimodal tasks without modifying the network architecture. SBNet involves extracting modality embeddings using modality-specific pre-trained networks and utilizing modality-invariant fully connected layers within a single branch to learn unified multimodal representations. However, it is not entirely input-agnostic, as it still necessitates separate processing of audio and visual data.

Inspired by the multimodal nature of human perception and the ability to make decisions based on any combination of available data, we introduce the first input-agnostic, multimodal, and unified biometric matching system. Our approach is designed to handle unimodal, multimodal, missing-modality, and cross-modal scenarios, supporting verification using audio, visual, and thermal modalities. Consequently, we overcome the limitations of existing works [12], [13], [14], [16], [17], [18], [19], [20], [21], [22], [23].

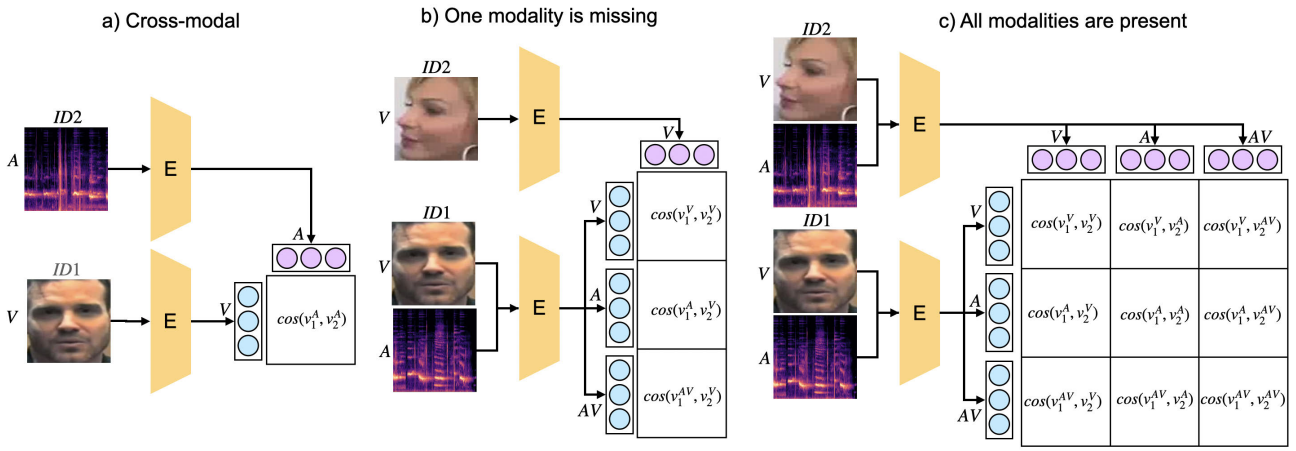

FIGURE 2. Biometric matching scenarios for our transformer-based unified system evaluated on the audio-visual VoxCeleb1 dataset: a) cross-modal, b) missing modality, and c) all available modalities are present.

The inclusion of the thermal modality is essential to ensure robust person verification, countering suboptimal environments that can adversely affect facial images and voice recordings, as well as potential deep-fakes in the visual and audio domains. The core of our system is a single transformer-based network with shared weights across all modalities. Our loss function aims to minimize discrepancies between different modalities, ensuring that the extracted embeddings of the same individual are closely clustered together, regardless of their original modality. Our main contributions are the following:

- An input-agnostic model that operates with audio, visual, and thermal data and performs multimodal, cross-modal, and unimodal biometric matching.
- A unified approach, where all learnable parameters are shared across all of the modalities.
- A multimodal prototypical network loss to map different modalities to the same representation space.
- Publicly available source code on our Github[1] to facilitate further research in this area.

The rest of this paper is organized as follows: Section II discusses the related works, focusing on the adaptations of transformers and prototypical networks for person recognition. Our unified transformer-based system, the multimodal prototypical network loss, and evaluation protocols are explained in Section III. Section IV presents and discusses the results. Finally, Section V concludes the paper and discusses future research directions.

II. RELATED WORKS

A. TRANSFORMERS FOR PERSON RECOGNITION

Transformers [28] have achieved state-of-the-art performance in various NLP tasks and are now widely used across multiple fields. Several works have adapted the transformer encoder for speaker recognition. Vaswani et al. explored five transformer variants and introduced a multi-view self-attention mechanism [28]. It balances capturing global dependencies and modeling local contexts by utilizing sliding windows of varying sizes for each attention head. In [29], local information modeling in the self-attention module is enhanced by restricting the attention context and introducing convolution operations. The SWIN transformer was adapted for the speaker verification task [30], processing input features at multiple scales using shifted local window self-attention to generate multi-scale output features.

All of these works [29], [30] are unimodal, operating only in the audio domain. The AV-SUPERB benchmark was the first attempt to unite audio and visual modalities for multiple tasks in speech and audio processing, including speaker recognition [31]. Meanwhile, the Vision Transformer (ViT) [32] has demonstrated an impressive capacity for generalization and transfer learning across various domains [33], [34], [35], [36]. Recognizing the substantial knowledge encoded in a pretrained ViT, our aim is to leverage it as a multimodal processor for understanding diverse modalities through their image representation and adapt it to the task of biometric matching.

B. PROTOTYPICAL NETWORKS

The prototypical networks were initially developed to tackle few-shot classification challenges, where labeled data is scarce [37], [38]. The core concept involves an embedding vector, or prototype, surrounded by data points of the same class. Classification is performed by computing the distances between queried examples and class prototypes, predicting classes based on these distances. The prototypes effectively represent the general characteristics of a class [37].

In the field of biometrics, a person's identity is traditionally verified based on a limited set of enrolled utterance samples, as the result prototypical networks have found success in audio-only speaker recognition [39], [40], [41]. Recently, this method has gained popularity for addressing different aspects of multimodal learning problems [42], [43], including action recognition [44], sound classification [45], speech

[1] https://github.com/IS2AI/unified_multimodal_transformer

recognition [46]. In our work, we adapt this loss to perform unimodal, multimodal and cross-modal matching by learning multimodal prototypes.

III. METHODS

We introduce a unified, single-branch and input-agnostic network to perform person verification. Figure 1 illustrates the key components of our architecture. Our system utilizes ViT [32] as the encoder to extract high-level feature representations from input data. The model can handle visual, thermal, and audio data, transforming them into a representation suitable for ViT. Thermal and visual data are provided as images, while raw audio undergoes conversion into log-mel spectrograms. These spectrograms, akin to image-like representations in the time-frequency domain, facilitate easy tokenization, similar to images.

All learnable parameters are shared across all of the three modalities. The encoder is fine-tuned for the person verification task using our multimodal prototypical network loss, explicitly designed to compare and contrast embeddings from different modality configurations. This approach enables our model to learn all possible combinations of modalities, offering adaptability to missing streams. As a result, the system is capable perform unimodal, multimodal, and cross-modal person verification (see Fig. 2).

A. THE ENCODER

The backbone of our verification system is the ViT pre-trained on ImageNet-21k, comprised of 14 million images and 21,843 classes [47]. The ViT's architecture closely follows the original transformer [28], with a few modifications at the input processing stage. Given an input image, the token embeddings are extracted by decomposing the input into a sequence of non-overlapping and fixed-sized patches (16×16 in this case), that are flattened and pushed through a learnable linear operator. The resulting sequence of embedded patches is prepended with a learnable class token and then combined with learnable positional embeddings to retain the position information of each patch.

The ViT utilizes only the encoder of the original transformer. The encoder comprises of a series of blocks with two main components. First, the series of embeddings go through a multiheaded self-attention (MSA) mechanism, which allows the ViT model to capture dependencies and relationships between different patches. Second, the output of MSA passes through a multilayer perceptron (MLP), which consists only of fully connected layers to capture complex and non-linear relationships in the data. MSA and MLP are both preceded by LayerNorm (LN) and followed by residual connections.

B. LOSS FORMULATION

The original loss was introduced in [37] to build prototypical networks for few-shot learning. Prototypical networks are trained in mini-batches. Each of them contains K classes, with a support set S and a query set Q for every class. A support is a labeled set of samples, that are used to predict classes of unlabeled samples, which are collectively called a query. In the context of biometric matching, a class is a person ID, and a sample is an utterance, that can be captured as an audio recording, or a facial image.

We denote $S_k = \{x_i, y_i\}$, $1 \leq i \leq N_S$, as the support set, where each x_i is an utterance of class k. The prototype of each class $c_k \in \mathbb{R}^D$ is the representative embedding of the class and calculated as the mean of embeddings in the support set:

$$c_k = \frac{1}{N_S} \sum_{i=1}^{N_S} E(x_i) \tag{1}$$

where E is an encoder that maps utterance data into the D-dimensional embedding space. In our case, it is the ViT encoder.

At the training stage, each query example $\{x_j, y_j\} \in Q$ is classified against K persons based on a softmax over Euclidean distances to each person's prototypes:

$$p(y_j = k|x_j) = \frac{\exp(-d(E(x_j), c_k))}{\sum_{k'=1}^{K} \exp(-d(E(x_j), c_{k'}))} \tag{2}$$

The prototypical network loss function (PNL) aims to minimize for each mini-batch the distance between a query feature vector and its true support prototype:

$$L_P = \sum_{k=1}^{K} \sum_{j=1}^{N_Q} -\log p(y_j = k|x_j) \tag{3}$$

As we adapt the loss to multimodal biometric matching, let us encode V, T, A as visual, thermal, and audio modalities, respectively. Thus, we further specify an utterance sample by adding the modality information as $x_i^u \in \mathbb{R}^{224\times224\times3}$, where $u \in \{A, V, T\}$. Note that each sample is a 3-channeled image of size 224×224, following the input requirement of the pretrained ViT encoder.

Since our evaluation scenarios consist of one-to-one comparisons, both the query and support sets are configured with a single utterance per person ID, following [40]. The utterances are represented by three modalities, which results in the following support and query sets for each class k:

$$S_k = \{x_{s,k}^A, x_{s,k}^V, x_{s,k}^T\}$$
$$Q_k = \{x_{q,k}^A, x_{q,k}^V, x_{q,k}^T\}$$

where $1 \leq q \leq N_q$, and $1 \leq s \leq N_s$.

Let us simplify the notation of the loss with introducing an embedding vector $v^u \in \mathbb{R}^{128}$ that is produced by the ViT encoder:

$$v_{s,k}^u = E(x_{s,k}^u)$$
$$v_{q,k}^u = E(x_{q,k}^u)$$

Noting that our system handles each data modality in a separate forward pass, the multimodal embedding vectors are

then computed based on the unimodal ones:

$$v_{q,k}^{VT} = \left(v_{q,k}^{V} + v_{q,k}^{T}\right)/2$$
$$v_{q,k}^{AV} = \left(v_{q,k}^{A} + v_{q,k}^{V}\right)/2$$
$$v_{q,k}^{AT} = \left(v_{q,k}^{A} + v_{q,k}^{T}\right)/2$$
$$v_{q,k}^{AVT} = \left(v_{q,k}^{A} + v_{q,k}^{V} + v_{q,k}^{T}\right)/3$$

In total, as illustrated in Fig. 1, the three data sources result in seven configurations of unimodal, bimodal and trimodal settings:

$$M = \{A, V, T, VT, AV, AT, AVT\}$$

The prototype of each configuration can be computed as:

$$c_k^{A} = v_{s,k}^{V}$$
$$c_k^{V} = v_{s,k}^{V}$$
$$c_k^{T} = v_{s,k}^{T}$$
$$c_k^{VT} = \left(v_{s,k}^{V} + v_{s,k}^{T}\right)/2$$
$$c_k^{AV} = \left(v_{s,k}^{A} + v_{s,k}^{V}\right)/2$$
$$c_k^{AT} = \left(v_{s,k}^{A} + v_{s,k}^{T}\right)/2$$
$$c_k^{AVT} = \left(v_{s,k}^{A} + v_{s,k}^{V} + v_{s,k}^{T}\right)/3$$

Since the distinction of query and support embeddings, allows us to specify the modality of the samples coming from different data sources, let us denote $m_s \in M$ and $m_q \in M$ as the modality configurations of the support and query sets respectively. We extend the definition of the prototypical network loss (3) to the multimodal case as:

$$L_P^{m_q,m_s} = -\sum_{k=1}^{K}\sum_{j=1}^{N_Q} \log \frac{\exp(-d(v_{j,k}^{m_q}, c_k^{m_s}))}{\sum_{k'=1}^{K}\exp(-d(v_{j,k'}^{m_q}, c_{k'}^{m_s}))} \quad (4)$$

The multimodal prototypical network network loss (MNPL) aggregates all possible combinations of modality settings:

$$L_{MP} = \sum_{m_q \in M}\sum_{m_s \in M} L_P^{m_q,m_s} \quad (5)$$

C. DATA PREPARATION

The model was trained and evaluated on two publicly available multimodal datasets suitable for biometric matching: SpeakingFaces [48] and VoxCeleb [49]. Statistics on both datasets are provided in Table 1. SpeakingFaces dataset was designed for biometric authentication and it consists of high-resolution thermal and visual spectral videos capturing fully framed faces, synchronized with audio recordings of individuals uttering commands in English. VoxCeleb is a large-scale human speech video dataset, that is widely recognized as the standard for benchmarking speaker recognition systems. It captures individuals speaking in various scenarios, including interviews and public speeches. Following the approach in [40], the VoxCeleb2 development split was deployed for training, and the VoxCeleb1 test split was used for evaluation, as detailed in Table 1. The data were reprocessed following our previous work [50], to extract facial regions from the visual and thermal images in both datasets.

TABLE 1. Statistics for the SpeakingFaces (SF) and VoxCeleb (VC) datasets.

#	SpeakingFaces			VoxCeleb	
	Train	Test	Valid	VC2 Dev	VC1 Test
speakers	100	20	22	5,994	40
utterances	9,307	1,843	1,886	1,092,009	4,874

Since we used the pretrained ViT encoder, we transformed all input data into 3-channeled images of size 224 × 224. For visual and thermal modalities, the extracted facial regions were first normalized for each modality separately and then resized to 224×224. For the raw audio data, we first extracted a random 3-second temporal segment with a sampling rate of 16 kHz and then transformed it into a 128-dimensional log-mel spectrograms. We increased the input channel of the audio spectrogram from 1 to 3, by duplicating the data, and then resized it to 224 × 224.

D. EVALUATION PROTOCOL

Due to the input-agnostic nature of our model, it can execute not only unimodal, multimodal, and cross-modal matching but also handle cases where one or more modalities are absent for a given subject. Figure 2 illustrates the capabilities of our unified system when trained on audio-visual input. In Fig. 2a, the cross-modal case is depicted, where a verification pair consists of a facial image for one speaker and an audio recording for the other, with the cosine similarity score between the corresponding extracted embedding vectors employed to quantify their similarity. Figure 2b showcases the scenario where complete audio-visual information is present for one speaker, and only one modality is available for the other. In this case, three scenarios are possible, depicted from top to bottom: the unimodal comparison (V versus V), the cross-modal matching from part (a), and the comparison of the fused audio-visual embedding with the visual one.

Figure 2c illustrates that when all data is available for both speakers, nine combinations can be executed to simulate all possible matching scenarios, including those presented in parts (a) and (b). These scenarios can be represented by a 3 × 3 matrix, with the main diagonal cells containing the comparisons with the same type of embeddings for both subjects. The top left cell represents the V versus V configuration, where only visual embeddings are compared with each other. The bottom right cell represents the audio-visual configuration, where the fused embeddings are contrasted with each other. The rest of the cells cover missing modality and cross-modal cases. The bottom left cell represents the comparison of audio-visual and visual embeddings (AV versus V), while the top center cell represents the cross-modal comparison of visual and audio embeddings (V versus A).

Using the same logic, the trimodal evaluation results in 49 combinations, as seven different embeddings can be constructed when all three modalities are present for a given speaker. This comprehensive approach illustrates the robustness and versatility of our model across various biometric matching scenarios.

E. IMPLEMENTATION DETAILS

The model implementation based on the PyTorch [51] framework was trained using an NVIDIA A100 graphics processing unit. Hyper-parameters were optimized separately for each modality setting with AdamW [52]. Each model was trained for 100 epochs and saved at each iteration. The models trained on SpeakingFaces, which performed the best on the validation set, were evaluated on the test set. All intermediate models trained on VoxCeleb2 development split were evaluated on the VoxCeleb1 test split. All experiments were repeated independently three times to minimize the effect of random initialization, and we report the mean of three experiments.

The models' performance was assessed based on the Equal Error Rate (EER) metric, a commonly employed measure in the evaluation of biometric matching [53]. The EER represents the point at which the false acceptance rate (FAR) and false rejection rate (FRR) are equal, offering a balanced assessment of the system's performance where both types of errors—incorrectly accepting a non-matching identity and incorrectly rejecting a matching identity—are minimized. A lower EER indicates superior overall system performance by minimizing both error types.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. SPEAKINGFACES RESULTS

We explored our approach on all possible combinations of modalities for the SpeakingFaces dataset. We trained three unimodal (visual, thermal and audio), three bimodal (visual-thermal, audio-visual, audio-thermal) and the tri-modal (audio-visual-thermal).

Figure 3 presents EER values on the validation and test sets for our unimodal and bimodal person verification systems. Figure 3a displays unimodal models denoted by V (visual), T (thermal), and A (audio). Each cell contains the EER on the validation (top) and test (bottom) sets for individual unimodal configurations.

Figures 3b-d collectively showcase evaluation scenarios of bimodal systems. They follow the evaluation logic outlined in Fig. 2c. In these, the featured systems are capable of producing up to three embeddings for a given subject (visual, thermal, and visual-thermal in Fig. 3b; visual, audio, and audio-visual in Fig. 3c; audio, thermal, and audio-thermal in Fig. 3d). For brevity, we will denote each embedding type with the first letter of its modality, e.g., AV for audio-visual and T for thermal. Since a person verification protocol involves comparing a pair of subjects, there could be in total nine comparison scenarios that can be represented by a 3×3 matrix. Each cell in the matrix stands for one of the scenarios. The main diagonal cells involve comparisons with the same type of embeddings for both subjects. For example, in Fig. 3c, the top left cell represents V versus V configuration, where only visual embeddings are compared with each other. The rest of the cells cover missing modality and cross-modal cases. For instance, the bottom left cell in the Fig. 3d represents the comparison of audio-thermal and thermal embeddings (AT versus T), while the top center cell represents the comparison of thermal and audio embeddings (T versus A). Similar to Fig. 3a, each cell lists the EER on the validation (top) and test (bottom) sets.

1) UNIMODAL SYSTEMS

In our previous study [50], we constructed encoders for unimodal thermal and visual systems using the default ResNet34 model [54]. For the unimodal audio system, we employed ResNet34 with self-attention pooling [55], instead of the global average pooling at the end of the residual network. In this work, while retaining the same input transformations, we utilize ViT as the encoder to capture all three modalities. ViT, having been pretrained on ImageNet-21k for visual image classification, unsurprisingly led to a reduction in EERs for visual verification on both evaluation sets: from 4.04% to 1.77% (validation) and from 4.10% to 2.4% (test) when compared to [50]. Overall, on both the validation and test sets, the visual modality demonstrates the lowest EER values among all unimodal configurations. The results for the unimodal thermal with ViT also showed improvement compared to the ResNet version, decreasing EER from 10.30% to 7.09% (validation) and from 10.86% vs 6.16% (test). The EERs for the unimodal audio models were comparable with our previously reported results, 10.82% versus 11.54% and 9.29% versus 9.94% for the validation and test sets, respectively. This suggests that ViT demonstrates strong results after fine-tuning on new modalities, adapting to thermal data and learning to interpret log-mel spectrograms as effectively as networks specifically designed for those modalities.

2) THE VISUAL-THERMAL SYSTEM

The joint training on the two modalities resulted in a substantial performance improvement in unimodal scenarios. Specifically, the EER improved from 1.77% to 0.68% (validation) and from 2.4% to 0.32% (test) in V versus V evaluation. A similar trend was observed in the T versus T scenario, with the EER improving from 7.09% to 3.55% (validation) and from 6.16% to 2.09% (test). The visual-thermal fusion configuration (VT versus VT) consistently outperforms individual unimodal configurations listed in Fig. 3a. On the test set, T versus VT and VT versus T (1.74% and 1.81%) outperform T versus T (2.09%). When thermal imagery is available for a pair of subjects, it's better to fuse at least one of them with visual information to improve the robustness. Furthermore, the visual-thermal evaluation

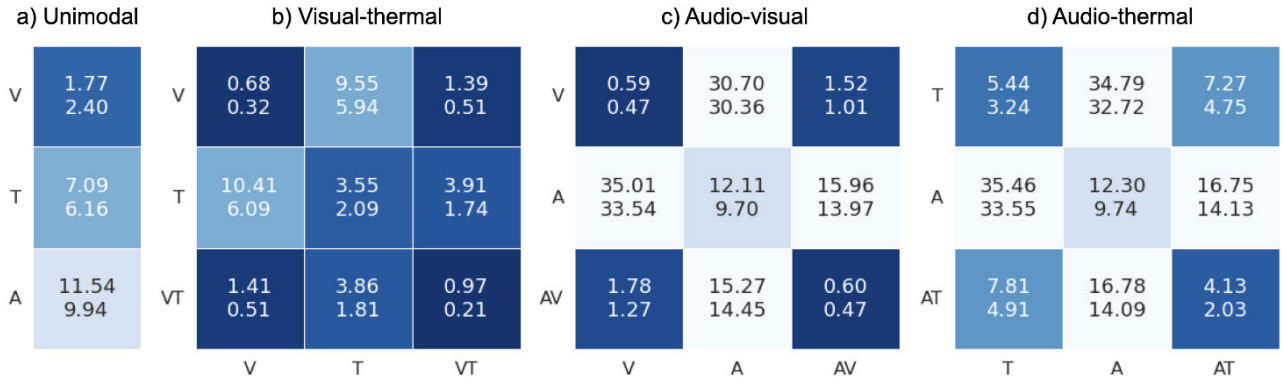

FIGURE 3. EER (%) of unimodal and bimodal models evaluated on the verification pairs of SpeakingFaces (SF) dataset. Each cell presents the EER (%) for the validation and test splits, displayed on the top and bottom rows, respectively. The reported values represent the mean results from repeated experiments. Part (a) shows the results for unimodal models. Parts (b) to (d) include evaluations for cross-modal and multimodal pairs. In these parts, the modalities on the horizontal axis correspond to the given modalities of person ID1, while those on the vertical axis correspond to the given modalities of person ID2. The darker the shade of the cells, the lower the EER (%), indicating better performance.

(VT versus VT) achieved EER of 0.21% on the test set and surpassed not only separate modality evaluations but also all of our previously reported results in [12]. The performance improvement indicates that visual and thermal modalities indeed complement each other. This supports the idea that combining multiple modalities can enhance the accuracy of person verification systems.

3) THE AUDIO-VISUAL SYSTEM

In evaluating the audio-only scenario (A versus A), comparable performance emerged between the unimodal audio model (11.54% validation and 9.94% test) and the audio-visual model (12.11% validation and 9.70% test). Joint training of the two modalities resulted in lower EERs for the visual-only scenario (V versus V). While the audio modality doesn't exhibit the same level of complementarity as thermal to the visual, it brought a notable drop in EER from 1.77% (unimodal visual) to 0.59% (audio-visual) in the validation set and from 2.4% (unimodal visual) to 0.47% (audio-visual) in the test set. Considering the inherent differences between audio and visual modalities, and their equal weight in computing the joint embedding, it is remarkable that the inclusion of audio information did not adversely impact the joint embedding and the training process. These findings underscore the robustness of the audio-visual model, demonstrating its capacity to integrate information from both modalities to enhance performance.

4) THE AUDIO-THERMAL SYSTEM

The evaluation of the audio-only scenario (A versus A) revealed comparable performance between the unimodal audio model (11.54% for validation and 9.94% test set) and the audio-thermal model (12.3% validation and 9.74% test set). Notably, joint training of both modalities yielded improvements when assessed in the thermal-only scenario (T versus T). The EER decreased from 7.09% (unimodal thermal) to 5.44% (audio-thermal) in validation and from 6.16% (unimodal thermal) to 3.24% (audio-thermal) in the test set. Although the reduction isn't as substantial as observed in the visual-thermal system, it's noticeable that training with the audio modality assisted in the thermal modality, despite significant differences in the appearance of the training data. Particularly interesting is the higher performance of the audio-thermal (AT versus AT) combination compared to the thermal-only scenario (T versus T). The fusion of audio and thermal data played a significant role in enhancing results for both modalities.

Overall, cross-modal scenarios generally result in the highest EER values for every bimodal system, highlighting that verifying subjects across different modalities presents more challenges compared to within-modal verification. In missing modality scenarios of audio-visual and audio-thermal models, the test and validation results suggest that the distribution of fused embeddings is closer to the embeddings extracted from facial images rather than ones from the log-mel spectrograms.

5) THE AUDIO-VISUAL-THERMAL SYSTEM

Figure 4 illustrates the performance of the trimodal model across both the validation and test splits. The trimodal model provides up to seven embeddings for each subject, including visual, thermal, audio, visual-thermal, audio-visual, audio-thermal, and audio-visual-thermal embedding vectors. In the context of a person verification protocol, where pairs of subjects are compared, this results in a matrix structure of 49 comparison scenarios, organized in a 7×7 matrix. Similar to Fig. 3, each cell in this matrix corresponds to a specific scenario, displaying the EER on the validation (top) and test (bottom) sets.

On the test set, the fusion of the three modalities achieves the EER of 0.27%, surpassing all other 48 configurations. This performance even surpasses our previously reported best EER of 2.48% for a trimodal network [12]. This reaffirms the efficacy of combining visual, thermal, and audio modalities

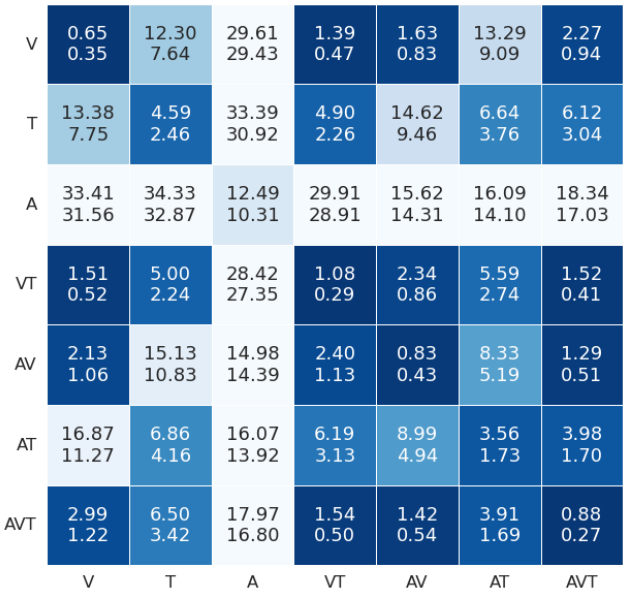

FIGURE 4. EER (%) of the trimodal model trained and evaluated on SpeakingFaces (SF) dataset. For each configuration (or cell), the EER (%) on the validation and test splits are presented at the top and bottom rows, respectively. We report the mean of the repeated experiments for unimodal, cross-modal and multimodal scenarios. The modalities on the horizontal axis correspond to the given modalities of person ID1, while those on the vertical axis correspond to the given modalities of person ID2. The darker the shade of the cells, the lower the EER (%).

within a person verification model, significantly enhancing its overall performance. At the same time, such a remarkable result was achieved by a singular network, that was able to capture three diverse modalities across the same set of shared weights.

The unimodal configurations maintain performance levels similar to their best-performing counterparts in the bimodal systems. EER values for audio-visual and visual-thermal configurations are comparable to their bimodal counterparts in Fig. 3 as well. However, the audio-thermal configuration outperforms its bimodal equivalent in Fig. 3, with EER values decreasing from 4.13% to 3.56% on the validation set and from 2.03% to 1.73% on the test set. This implies that training with visual modality has enriched the model, enabling it to perform well even in the audio-thermal configuration where visual information was not initially required.

In line with the results presented in Fig. 3, two notable observations highlight the complementarity of thermal and visual modalities for the trimodal model. First, the fusion of visual and thermal information is stronger than each individual modality. On the test set, visual-thermal configuration (VT versus VT) achieves 0.29%, surpassing the visual-only (V versus V) and thermal-only (T versus T) configurations, which reach 0.35% and 2.46%, respectively. Second, if visual information is available for at least one subject in a pair, incorporating it with the thermal information enhances the performance of the verification process. The evaluation scenarios, T versus VT and VT versus T demonstrate a lower EER value than T versus T, with 2.26% and 2.24%, respectively, being lower than 2.46% in the T versus T case.

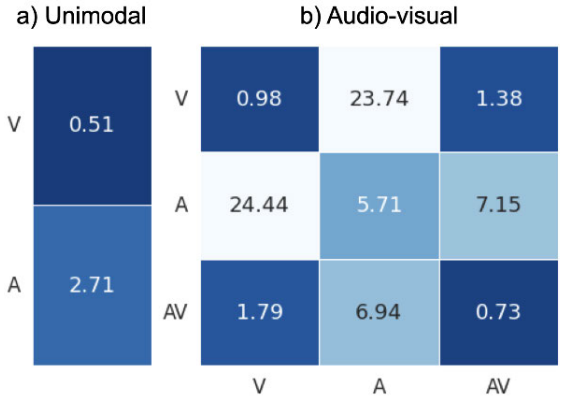

FIGURE 5. EER (%) of a) unimodal audio and unimodal visual, and b) audio-visual models trained on VoxCeleb2 dev split and evaluated on VoxCeleb1 (VC1) test split. We report the mean of the repeated experiments. Each cell represents the EER (%) for the given modality pair. The modalities on the horizontal axis correspond to the given modalities of person ID1, while those on the vertical axis correspond to the given modalities of person ID2. The darker the shade of the cells, the lower the EER (%), indicating better performance.

TABLE 2. Unimodal verification EER (%) on VoxCeleb1 (VC1) test split of unimodal models trained on VoxCeleb2 development split. For our models, the three EER values represent results obtained using three different random seeds, which were consistently used across all train configurations.

Test condition	VC1 Test
Unimodal visual of [14]	3.9
Unimodal visual of [23]	3.04
Unimodal visual of [12]	3.87 4.20 3.73
Unimodal visual (ours)	0.51 **0.49** 0.51
Unimodal audio of [14]	2.2
Unimodal audio of [23]	1.63
Unimodal audio of [12]	2.21 2.04 2.00
Unimodal audio of [56]	2.56
Unimodal audio of [29]	1.96
Unimodal audio of [57]	1.34
Unimodal audio (ours)	2.70 2.93 **2.51**

To sum up, the SpeakingFaces results verify the strength of a unified transformer network to tackle various modalities for the person verification task.

B. VOXCELEB RESULTS

Figure 5 showcases the performance of the unimodal audio, unimodal visual, and audio-visual models trained on the VoxCeleb2 development set and evaluated on the VoxCeleb1 test set. In Fig. 5a, we present the performance of the systems trained exclusively on single-modality data. We compare these results with state-of-the-art models listed in Table 2. Notably, our ViT-based approach outperforms all others in the visual domain. While our unimodal audio model exhibits slightly lower performance compared to the counterparts, it's important to note that these comparisons involve systems with encoders custom-built for audio data. In contrast, our ViT encoder maintains the same structure for both modalities, emphasizing its versatility and effectiveness.

TABLE 3. Unimodal verification EER (%) on VoxCeleb1 (VC1) test split of the models capable of both unimodal and cross-modal cases. For our models, the three EER values represent results obtained using three different random seeds, which were consistently used across all train configurations.

Test condition	VC1 Test
Visual-only of [14]	3.7
Visual-only of [21]	13.1
Visual-only (our audio-visual model)	1.03 1.09 **0.83**
Audio-only of [14]	6.1
Audio-only of [21]	9.6
Audio-only (our audio-visual model)	5.76 6.40 **4.97**

TABLE 4. Cross-modal verification EER (%) on VoxCeleb1 (VC1) test split from previous studies and our unified transformer.

Method	VC1 Test
Learnable Pins [16]	29.6
Deep Latent Space [17]	29.5
Multi-view Approach [14]	28.0
Disentangled Representation Learning [20]	25.0
Single-branch (Git) [21]	25.7
Unified transformer (our audio-visual)	**24.1**

In Fig. 5b, we focus on the evaluation of the audio-visual model across the nine scenarios, featured in Fig. 3c. The fusion of audio and visual modalities demonstrates superior performance compared to each individual configuration. That is, visual-only (V versus V) yields a mean EER of 0.98%, audio-only (A versus A) gives a mean EER of 5.71%, while audio-visual (AV versus AV) achieves 0.73%. However, it is worth noting that audio-visual, although strong, is worse than the unimodal visual model featured in Fig. 5a, which exhibits an EER of 0.51%. Additionally, the visual-only configuration is not as strong as the unimodal visual model, with the EER decreasing from 0.98% to 0.51%. A similar situation is observed with the audio-only configuration, transitioning from 5.71% to 2.71%.

Our audio-visual system stands out compared to state-of-the-art models capable of simultaneously handling unimodal, multimodal, and cross-modal verification on the VoxCeleb1 test set. Table 3 provides a comparison of our bimodal approach in single-modality scenarios with the Multi-view Approach [14] and the Single-branch (Git) [21]. In both visual-only and audio-only cases, our ViT-based system significantly outperforms its counterparts.

Table 4 contrasts the performance of our audio-visual model with state-of-the-art models specifically in cross-modal verification scenarios. Cross-modal verification poses a significant challenge, as it involves matching entirely different data types. The VoxCeleb1 test and VoxCeleb2 development splits are disjoint, adding to the complexity, as we aim to associate the face of a previously unseen individual with the voice of someone previously unheard during training. We report the mean of cross-modal evaluations (A versus V, and V versus A). In this context, our model exhibits superior performance when compared to Learnable Pins [16], Deep Latent Space [17], Multi-view Approach [14], DIMNet [22], Disentangled Representation Learning [20], and Single-branch (Git) [21].

In summary, the performance on the VoxCeleb dataset also underscores the efficacy of our unified approach. The ViT-based encoder, shared across the two modalities and coupled with the multimodal prototypical network loss, demonstrates proficiency in handling bimodal data. The model effectively matches diverse combinations of modalities, proving its adaptability to cross-modal scenarios.

V. CONCLUSION AND FUTURE WORKS

In this study, we introduced a transformer model for biometric verification and demonstrated its versatility across modalities. Joint training on audio, visual, and thermal data using a unified ViT with multimodal prototypical network loss led to significant improvements in EER. Unimodal models trained and tested on the SpeakingFaces dataset showed enhanced performance in visual and thermal modalities compared to our previous ResNet-based approach, with comparable results for the audio modality. While ViT showcased adaptability to thermal and audio modalities, further enhancements are needed to strengthen its performance with audio compared to custom networks tailored for this type of data. The bimodal training further decreased the EER on both single- and multi-modal scenarios of bimodal models. Overall, the bimodal configurations consistently outperformed the unimodal models. The trimodal single network was able to take the best of its unimodal and bimodal counterparts and achieved the EER of 0.27% on the test set and 0.88% on the validation set, surpassing our previously reported ResNet-based trimodal model.

To ensure a fair benchmarking of our unified transformer on the VoxCeleb1 test set, we assessed its performance against other models designed for both unimodal and cross-modal verification. Our bimodal system outperformed the state-of-the-art results in both scenarios. The results of the unimodal models confirmed ViT's suitability for handling facial data, outperforming other visual methods. However, improvements are needed in handling raw audio data to enhance competitiveness in this domain.

The presented framework involved a simple fusion strategy, providing equal weight to each modality in computing multimodal embeddings. In future work, we plan to explore advanced attention mechanisms for more effective fusion. By integrating attention mechanisms that are jointly trained with the encoder, we aim to dynamically adjust the contributions of each modality based on the context, thereby enhancing the overall fusion process and improving performance.

Overall, the achieved improvements of the trimodal model on the SpeakingFaces dataset and the outstanding performance of the audio-visual model on the VoxCeleb dataset, highlight the efficacy of our unified multimodal approach in tackling a variety of scenarios for biometric checking.

REFERENCES

[1] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometrics recognition using deep learning: A survey," *Artif. Intell. Rev.*, vol. 56, no. 8, pp. 8647–8695, Aug. 2023.

[2] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *Comput. Vis. Image Understand.*, vol. 189, Dec. 2019, Art. no. 102805.

[3] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Netw.*, vol. 140, pp. 65–99, Aug. 2021.

[4] C. Militello, L. Rundo, S. Vitabile, and V. Conti, "Fingerprint classification based on deep learning approaches: Experimental findings and comparisons," *Symmetry*, vol. 13, no. 5, p. 750, Apr. 2021.

[5] G. Liu, W. Zhou, L. Tian, W. Liu, Y. Liu, and H. Xu, "An efficient and accurate iris recognition algorithm based on a novel condensed 2-ch deep convolutional neural network," *Sensors*, vol. 21, no. 11, p. 3721, May 2021.

[6] E. Maiorana, "Deep learning for EEG-based biometric recognition," *Neurocomputing*, vol. 410, pp. 374–386, Oct. 2020.

[7] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, "Face recognition systems: A survey," *Sensors*, vol. 20, no. 2, p. 342, Jan. 2020.

[8] M. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality adaptive margin for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18729–18738.

[9] 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.

[10] J. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1–16.

[11] S. Hörmann, A. Moiz, M. Knoche, and G. Rigoll, "Attention fusion for audio-visual person verification using multi-scale features," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Nov. 2020, pp. 281–285.

[12] M. Abdrakhmanova, T. Unaspekov, and H. Atakan Varol, "Multimodal person verification with generative thermal data augmentation," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 6, no. 1, pp. 43–53, Jan. 2024.

[13] X. Jing, L. He, Z. Song, and S. Wang, "Audio–visual fusion based on interactive attention for person verification," *Sensors*, vol. 23, no. 24, p. 9845, Dec. 2023.

[14] L. Sari, K. Singh, J. Zhou, L. Torresani, N. Singhal, and Y. Saraf, "A multi-view approach to audio-visual speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6194–6198.

[15] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8427–8436.

[16] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable PINs: Cross-modal embeddings for person identity," *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 71–88.

[17] S. Nawaz, M. K. Janjua, I. Gallo, A. Mahmood, and A. Calefati, "Deep latent space learning for cross-modal mapping of audio and visual signals," in *Proc. Digit. Image Comput., Techn. Appl.*, Dec. 2019, pp. 1–7.

[18] R. Tao, R. K. Das, and H. Li, "Audio-visual speaker recognition with a cross-modal discriminative network," in *Proc. Interspeech*, Oct. 2020, pp. 2242–2246.

[19] A. Nagrani, J. S. Chung, S. Albanie, and A. Zisserman, "Disentangled speech embeddings using cross-modal self-supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6829–6833.

[20] H. Ning, X. Zheng, X. Lu, and Y. Yuan, "Disentangled representation learning for cross-modal biometric matching," *IEEE Trans. Multimedia*, vol. 24, pp. 1763–1774, 2022.

[21] M. S. Saeed, S. Nawaz, M. H. Khan, M. Z. Zaheer, K. Nandakumar, M. H. Yousaf, and A. Mahmood, "Single-branch network for multimodal training," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[22] Y. Wen, M. Al Ismail, W. Liu, B. Raj, and R. Singh, "Disjoint mapping network for cross-modal matching of voices and faces," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–19.

[23] Y. Qian, Z. Chen, and S. Wang, "Audio-visual deep neural network for robust person verification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, no. 1, pp. 1079–1092, Jun. 2021.

[24] L. Kezebou, V. Oludare, K. Panetta, and S. Agaian, "TR-GAN: Thermal to RGB face synthesis with generative adversarial network for cross-modal face recognition," in *Mobile Multimedia/Image Processing, Security, and Applications*, vol. 11399. Bellingham, WA, USA: SPIE, 2020, pp. 158–168.

[25] M. Kowalski, A. Grudzien, and K. Mierzejewski, "Thermal–visible face recognition based on CNN features and triple triplet configuration for on-the-move identity verification," *Sensors*, vol. 22, no. 13, p. 5012, Jul. 2022.

[26] Y. Gavini, A. Agarwal, and B. M. Mehtre, "Thermal to visual person re-identification using collaborative metric learning based on maximum margin matrix factorization," *Pattern Recognit.*, vol. 134, Feb. 2023, Art. no. 109069.

[27] L. Sari, S. Thomas, and M. Hasegawa-Johnson, "Training spoken language understanding systems with non-parallel speech and text," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 8109–8113.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–30.

[29] B. Han, Z. Chen, and Y. Qian, "Local information modeling with self-attention for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6727–6731.

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[31] Y. Tseng, L. Berry, Y.-T. Chen, I.-H. Chiu, H.-H. Lin, M. Liu, P. Peng, Y.-J. Shih, H.-Y. Wang, H. Wu, P.-Y. Huang, C.-M. Lai, S.-W. Li, D. Harwath, Y. Tsao, A. Mohamed, C.-L. Feng, and H.-Y. Lee, "AV-SUPERB: A multi-task evaluation benchmark for audio-visual representation models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 6890–6894.

[32] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai, "An image is worth 16×16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929*.

[33] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-supervised audio spectrogram transformer," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 10, pp. 10699–10709, doi: 10.1609/aaai.v36i10.21315.

[34] R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra, "Omnivore: A single model for many visual modalities," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16081–16091.

[35] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. Vasudev Alwala, A. Joulin, and I. Misra, "ImageBind one embedding space to bind them all," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15180–15190.

[36] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[37] J. Wang and Y. Zhai, "Prototypical Siamese networks for few-shot learning," in *Proc. IEEE 10th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2020, pp. 178–181.

[38] Z. Ji, X. Chai, Y. Yu, Y. Pang, and Z. Zhang, "Improved prototypical networks for few-shot learning," *Pattern Recognit. Lett.*, vol. 140, pp. 81–87, Dec. 2020.

[39] J. Wang, K.-C. Wang, M. T. Law, F. Rudzicz, and M. Brudno, "Centroid-based deep metric learning for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3652–3656.

[40] J. S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Proc. Interspeech*, 2020, pp. 2977–2981.

[41] T. Ko, Y. Chen, and Q. Li, "Prototypical networks for small footprint text-independent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6804–6808.

[42] Y. Fan, W. Xu, H. Wang, J. Wang, and S. Guo, "PMR: Prototypical modal rebalance for multimodal learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 20029–20038.

[43] F. Pahde, M. Puscas, T. Klein, and M. Nabi, "Multimodal prototypical networks for few-shot learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2643–2652.
[44] Y. Wanyan, X. Yang, C. Chen, and C. Xu, "Active exploration of multimodal complementarity for few-shot action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6492–6502.
[45] S. S. Kushwaha and M. Fuentes, "A multimodal prototypical approach for unsupervised sound classification," in *Proc. Interspeech*, Aug. 2023, pp. 266–270.
[46] Y.-K. Zhang, D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "Audio-visual generalized few-shot learning with prototype-based co-adaptation," in *Proc. Interspeech*, Sep. 2022, pp. 531–535.
[47] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik, "Imagenet-21k pretraining for the masses," in *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, vol. 1, 2021, pp. 1–20.
[48] M. Abdrakhmanova, A. Kuzdeuov, S. Jarju, Y. Khassanov, M. Lewis, and H. A. Varol, "SpeakingFaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams," *Sensors*, vol. 21, no. 10, p. 3465, May 2021.
[49] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, Aug. 2017, pp. 1–17.
[50] M. Abdrakhmanova, S. Abushakimova, Y. Khassanov, and H. A. Varol, "A study of multimodal person verification using audio-visual-thermal data," in *Proc. Speaker Lang. Recognit. Workshop*, Jun. 2022, pp. 233–239.
[51] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–23.
[52] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–20.
[53] H. Gish, "Elements of speaker verification," in *Biometric Systems*. Cham, Switzerland: Springer, 2005, pp. 115–136.
[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[55] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," in *Proc. Speaker Lang. Recognit. Workshop*, Jun. 2018, pp. 1–27.
[56] R. Wang, J. Ao, L. Zhou, S. Liu, Z. Wei, T. Ko, Q. Li, and Y. Zhang, "Multi-view self-attention based transformer for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6732–6736.
[57] M. Sang, Y. Zhao, G. Liu, J. H. L. Hansen, and J. Wu, "Improving transformer-based networks with locality for automatic speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

MADINA ABDRAKHMANOVA (Member, IEEE) received the B.S. degree in computer science from the University of Illinois at Urbana–Champaign, USA, in 2014, and the M.S. degree in computer science from the University of California at Irvine, Irvine, in 2019. In 2014, she joined Nazarbayev University, where she is currently a Data Scientist with the Institute of Smart Systems and Artificial Intelligence. Her research interests include computer vision, machine learning, multimodal learning, and generative models.

ASSEL YERMEKOVA received the B.S. degree in nuclear physics from Eurasian National University, Astana, Kazakhstan, in 2019, and the M.S. degree in computer science from Skolkovo Institute of Science and Technology, Moscow, in 2020. From 2022 to 2023, she was a Research Assistant with the Institute of Smart Systems and Artificial Intelligence, Nazarbayev University. Her research interests include deep learning, speech processing, multimodal, and generative models.

YULIYA BARKO is currently pursuing the B.S. degree in computer science with Nazarbayev University, Astana, Kazakhstan. She is a Research Assistant with the Institute of Smart Systems and Artificial Intelligence, Nazarbayev University. Her research interests include deep learning, multimodal learning, self-supervised learning, and computer vision.

VLADISLAV RYSPAYEV received the B.S. degree in mathematics from Nazarbayev University, Astana, Kazakhstan, in 2024. He is currently a Research Assistant with the Institute of Smart Systems and Artificial Intelligence, Nazarbayev University. His research interests include federated learning, distributed learning, and cross-modal learning.

MEDET JUMADILDAYEV is currently pursuing the B.S. degree in computer science with Nazarbayev University, Astana, Kazakhstan. He is a Research Assistant with the Institute of Smart Systems and Artificial Intelligence, Nazarbayev University. His research interests include deep learning, multimodal learning, and natural language processing.

HUSEYIN ATAKAN VAROL (Senior Member, IEEE) received the B.S. degree in mechatronics engineering from Sabanci University, Istanbul, Türkiye, in 2005, and the M.S. and Ph.D. degrees in electrical engineering from Vanderbilt University, Nashville, TN, USA, in 2007 and 2009, respectively.

From 2009 to 2011, he was a first Postdoctoral Research Associate and a Research Assistant Professor with the Center for Intelligent Mechatronics, Department of Mechanical Engineering, Vanderbilt University. In 2011, he joined the Faculty of Nazarbayev University, Astana, Kazakhstan, where he is currently a Full Professor of robotics and directs with the Institute of Smart Systems and Artificial Intelligence (ISSAI). He has published more than 120 technical papers on these topics in reputable international journals and conferences. He holds multiple international patents in the area of bionics. He has attracted multiple competitively-funded research grants from the Ministry of Education and Science of Kazakhstan, World Bank, and Nazarbayev University Research Fund. His research interests include artificial intelligence, biometrics, machine learning, soft robotics, sensor fusion, and tensegrity. He served as a technical and an associate editor for prominent journals and conferences.