Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061187

ARTICLE

Cross-Modal Simplex Center Learning for Speech-Face Association

Qiming Ma, Fanliang Bu*, Rong Wang, Lingbin Bu, Yifan Wang and Zhiyuan Li

School of Information Network Security, People's Public Security University of China, Beijing, 100038, China

*Corresponding Author: Fanliang Bu. Email: bufanliang@sina.com

Received: 19 November 2024; Accepted: 25 December 2024; Published: 06 March 2025

ABSTRACT: Speech-face association aims to achieve identity matching between facial images and voice segments by aligning cross-modal features. Existing research primarily focuses on learning shared-space representations and computing one-to-one similarities between cross-modal sample pairs to establish their correlation. However, these approaches do not fully account for intra-class variations between the modalities or the many-to-many relationships among cross-modal samples, which are crucial for robust association modeling. To address these challenges, we propose a novel framework that leverages global information to align voice and face embeddings while effectively correlating identity information embedded in both modalities. First, we jointly pre-train face recognition and speaker recognition networks to encode discriminative features from facial images and voice segments. This shared pre-training step ensures the extraction of complementary identity information across modalities. Subsequently, we introduce a cross-modal simplex center loss, which aligns samples with identity centers located at the vertices of a regular simplex inscribed on a hypersphere. This design enforces an equidistant and balanced distribution of identity embeddings, reducing intra-class variations. Furthermore, we employ an improved triplet center loss that emphasizes hard sample mining and optimizes inter-class separability, enhancing the model's ability to generalize across challenging scenarios. Extensive experiments validate the effectiveness of our framework, demonstrating superior performance across various speech-face association tasks, including matching, verification, and retrieval. Notably, in the challenging gender-constrained matching task, our method achieves a remarkable accuracy of 79.22%, significantly outperforming existing approaches. These results highlight the potential of the proposed framework to advance the state of the art in cross-modal identity association.

KEYWORDS: Speech-face association; cross-modal learning; cross-modal matching; cross-modal retrieval

1 Introduction

Speech data not only includes physical attributes such as pitch, intensity, duration, and timbre but is also closely tied to the physiological characteristics of the vocal apparatus. In contrast, facial data provides abundant structural and visual information. The latent features in both speech and facial data can reflect attributes such as gender, ethnicity, and age [1,2]. Capturing these latent features has enabled remarkable advancements in face recognition and speaker identification, and these features similarly contribute to cross-modal matching and retrieval tasks. Previous studies have identified a correlation between voice and facial characteristics [3], vocal features such as intonation, timbre, and rhythm can reflect a speaker's vocal anatomy, geographic background, and race, which are closely tied to facial features, enabling individuals to form cognitive impressions of facial appearance even without visual cues. Recent neurological research has further confirmed that humans have the innate ability to match faces with voices they hear [4]. With the growing influence of artificial intelligence, researchers have begun employing deep learning algorithms to automate the correlation between these two modalities [5–8]. This technology has wide-ranging applications

in areas such as criminal investigation, video conferencing, and entertainment, where information from one modality can be used to synchronize with the other in real-time, enhancing both system intelligence and the interactive experience.

The currently prevalent approach employs a two-stream network architecture (as illustrated in Fig. 1), where encoders are first established for both face and voice modalities to directly map them into a common feature space. Next, loss functions such as cross-entropy or mean squared error are applied to maximize inter-class differences in the label space of the samples. Finally, a cross-modal matching objective function is designed to align images and voices within the joint embedding space. These methods typically utilize contrastive loss [5], N-pair loss [7] and triplet loss [8] as objective functions. Under supervised conditions, triplet loss has demonstrated superior stability and performance in enhancing feature discriminability, outperforming methods based on contrastive loss and N-pair loss [9].

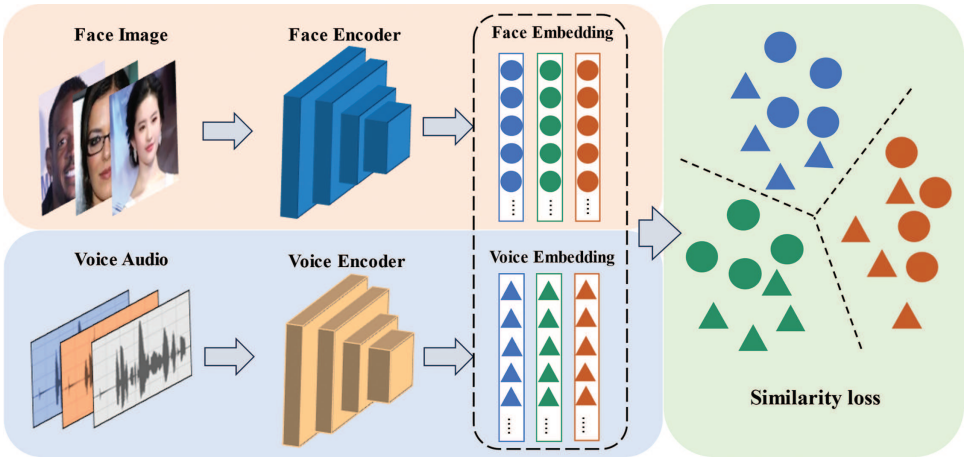

Figure 1: Two-stream framework for speech-face cross-modal association

However, a significant semantic gap exists between different modalities. Specifically, facial images for each identity display diverse poses, lighting conditions, and styles, while speech varies greatly in content and emotion. These differences result in substantial distributional discrepancies between facial and speech features when mapped to the latent space, leading to two key issues: (a) First, prior studies often align facial and speech features directly in the latent space without incorporating a global cross-modal interaction process. This omission can result in pronounced intra-modal gaps, limiting the ability to effectively model correlations between the two modalities. (b) Second, most methods compute similarity loss only for cross-modal sample pairs, relying on one-to-one distance quantification at the model's output. This approach overlooks the complex many-to-many relationships between speech segments and facial images, hindering the full exploration of potential connections across modalities.

To address the aforementioned challenges, this paper proposes a speech-face association learning method based on cross-modal simplex center loss. Unlike traditional center loss methods, we select identity centers from the vertices of a regular simplex inscribed within a hypersphere instead of relying on learnable parameters. This approach ensures that the identity centers have equal Euclidean and cosine distances, facilitating the optimization of inter-class relationships and improving the stability and performance of the model. However, due to the involvement of both speech and face modalities, directly minimizing the distances between the features of these two modalities and their respective class centers may lead to poor generalization. To overcome this limitation, we introduce a cross-modal projection mechanism during

center alignment to promote intra-class consistency across modalities. Additionally, to enhance the model's ability to distinguish challenging samples, we incorporate an improved **triplet center loss**. This loss function effectively identifies hard negative samples within the dataset, enabling the model to differentiate between samples belonging to different identities that appear similar in the feature space. This further strengthens the understanding of the relationships between speech and facial features. The experimental results demonstrate that our proposed method outperforms recent speech-face association approaches.

The primary contributions of this work can be summarized as follows:

- We propose a novel **cross-modal simplex center loss**, which is easier to optimize and enhances inter-class relationships while providing improved stability compared to traditional center loss methods.
- We introduce a new framework for cross-modal retrieval tasks that combines joint pretraining, cross-modal simplex center loss, and an improved triplet center loss, making it extensible to various supervised cross-modal retrieval tasks.
- We validate the effectiveness of the proposed method through experiments on speech-face cross-modal matching, verification, and retrieval tasks under different settings.

2 Related Work

In this section, we first introduce the task details and common solutions for cross-modal retrieval. Next, we analyze the existing methods for speech-face cross-modal association learning and identify their limitations. Based on these analyses, we propose our improvements to address these shortcomings.

2.1 Cross-Modal Retrieval

Cross-modal learning facilitates the transfer of information between data from different sensory modalities, such as images, speech, and text. This approach explores the correspondence between samples by leveraging the correlation or complementarity among different modalities and has widespread applications in fields such as transformation, generation, and retrieval [10–12]. Among these applications, cross-modal retrieval methods primarily consist of two stages: feature extraction and feature alignment [13].

In the feature extraction stage, there are primarily two approaches. The first approach involves encoding each modality independently by designing separate encoders tailored to the characteristics of each modality. Recent methods utilize pre-training on large-scale datasets to extract deep semantic information from different modalities [14–16]. Additionally, some approaches adapt the encoding architecture based on the characteristics of the modality data, employing LSTM, Transformer, or other architectures for encoding sequential data such as video, speech, and text, while using convolutional neural networks for image encoding [17]. In contrast, the second approach adopts interactive learning methods. To leverage inter-modal correlation information during encoding, You et al. [18] proposed an improved CLIP with parameter-sharing encoders across different modalities, where certain modules or layers of the model are modality-shared, thus narrowing the semantic similarity gap between modalities. Jiang et al. [19] constructed a multi-modal interactive encoder that employs self-attention and cross-attention mechanisms to establish associations between vision and text.

The feature alignment stage primarily encompasses two approaches. The first approach is joint embedding space learning, where cross-modal data are projected into a shared latent space for direct comparison. To address the differences between modalities, Zhang and Lu [20] employ a cross-modal projection method that classifies the vector projections of representations from one modality to another during feature alignment, thereby enhancing the compactness of features for each category. Peng and Qi [21] introduce adversarial training to eliminate differences between modalities, generating modality-invariant features within the

common space. The second approach is pairwise similarity learning, which focuses on developing a similarity function that ensures the similarity of cross-modal sample pairs with the same label is greater than that of negative pairs. Many cross-modal retrieval studies optimize the network model by maximizing the similarity of cross-modal sample pairs [17,19,22]. Unlike works that only consider matched pairs, Yan et al. [23] separately learn single-modal and cross-modal proxies during the feature encoding process, effectively capturing cross-modal similarities between samples. Liu et al. [24] adopt vector quantization methods to learn shared discrete embeddings for cross-modal samples with the same label in the latent space. Jing et al. [25] extend center loss into a cross-modal version, learning center vectors for each category in the joint latent space and minimizing the distance between data, such as 2D images, 3D point clouds, and meshes, and their corresponding category center vectors.

In audio-visual cross-modal retrieval tasks, Surís et al. [26] employ a two-stream network with stacked fully connected layers. The network maps videos and audio to a joint embedding space, achieving audio-visual retrieval by optimizing the Euclidean distance and cosine similarity between video-audio cross-modal samples. Gabeur et al. [27] introduce a pre-training method to learn video representations, alternating the use of RGB, audio, and transcribed speech as supervision during training. Unlike the aforementioned methods that consider only pairwise similarity, Hao et al. [28] develop a common attention module to interactively process video and audio inputs, thereby better uncovering the semantic relationships between them. Yuan et al. [29] first perform cross-modal fusion on the raw data and then adopt an Attention-Fuse-then-Separate strategy to implicitly capture cross-modal dependencies and common representations. The development of these studies highlights the potential for machines to associate visual events with sound signals. This paper aims to explore the more specific task of face-speech biometric matching.

2.2 Speech-Face Association

Unlike most previous audio-visual cross-modal retrieval tasks, speech-face association primarily focuses on identity for feature alignment. It aims to investigate whether potential attributes—such as gender, age, race, pronunciation structure, and language habits—contained in faces and voices correspond to one another. Kim et al. [8] have experimentally demonstrated that humans can match unfamiliar voices and faces with an accuracy exceeding chance, indicating that automatic association through deep neural networks holds great potential. Current research treats the matching problem as a binary classification task [6], where samples from one modality serve as queries to predict positive cases with the same identity in the other modality. Nagrani et al. [5] integrate cross-modal verification and retrieval into the speech-face association task, establishing it as the mainstream evaluation standard for this field. Most studies map speech and face features to the same dimensional latent space using two-stream networks, followed by deep metric learning methods for feature alignment, such as contrastive loss, N-pair loss and triplet loss. In supervised cross-modal learning, Nawaz et al. [30] and Wen et al. [31] have shown that optimizing feature discriminability with a modality-shared ID classifier can implicitly achieve modality alignment. Inspired by existing cross-modal retrieval methods, Wen et al. [32] introduce common variables of speech and face modalities, such as gender and age information, as supervision to establish a connection between voice and face. Zheng et al. [33] propose a modality discriminator-based method that reduces the discrepancy between the two modalities in the latent space through adversarial learning.

Although existing methods have attempted to address the heterogeneity between modalities from various perspectives, such as employing extensive manual annotations or designing complex encoding networks, they often face challenges in fully leveraging global cross-modal interactions and addressing the intricate many-to-many relationships inherent in cross-modal data. In this work, we propose a joint pre-training strategy that leverages identity supervision, aligning shared identity features during the feature

encoding stage to bridge modality differences. Unlike traditional similarity functions, our proposed cross-modal simplex center loss not only captures the many-to-many relationships between speech segments and facial images but also learns highly discriminative representations, thereby improving cross-modal matching and retrieval performance.

3 Methods

Our goal is to align the modality-independent identity information contained in voice segments and facial images, thereby enabling a range of tasks, including face-speech/speech-face matching, verification, and retrieval. The framework of the proposed method is illustrated in Fig. 2. First, for feature extraction, we jointly pre-train face recognition and speaker recognition networks as encoders for images and voice segments. After extracting face and speech embeddings, we introduce a new alignment strategy based on cross-modal simplex center loss in Section 3.2 (Fig. 2b). This strategy encourages face and speech features to align in a many-to-many manner according to identity by minimizing the distance between the mutual projections of the two modalities and the identity center vectors corresponding to the vertices of the regular simplex. To address the common issue of similar samples that do not belong to the same identity in large-scale face and speech open sets, we present cross-modal triplet center loss in Section 3.3 (Fig. 2c), which mines hard negative samples in the mini-batch as an effective supplement to center alignment.

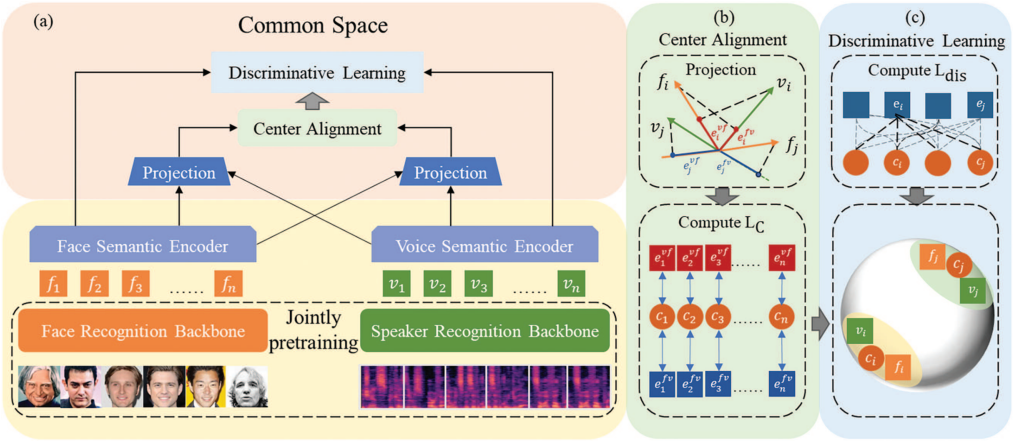

Figure 2: Proposed speech-face association framework based on cross-modal simplex center learning

3.1 Joint Pre-Training

Previous work has demonstrated that identity classification in a single modality facilitates cross-modal associations between speech and facial images. Based on this finding, this work utilizes identity as supervisory information to jointly pre-train face recognition and speaker recognition tasks on the training set. The pre-trained networks are then used to extract feature embeddings for facial images and speech segments, which are subsequently aligned through further feature alignment learning. To validate the effectiveness of the proposed joint pre-training strategy, all experiments are conducted using baseline networks.

For face recognition, we utilize VGG-Face [34], which accepts 112 × 112 × 3 RGB facial images as input. Data augmentation strategies, including random rotation and random cropping, are employed during pre-training. Two stacked fully connected layers are added at the end of the network, mapping the output dimensions to a 128-dimensional speech-face common space as the face image representation. For speaker

recognition, we employ VGGish [35]. Each voice segment is first divided into non-overlapping 960 ms frames, from which one frame is randomly selected for Short-Time Fourier Transform (STFT), resulting in a 96 × 64 Log Mel Spectrogram. During the extraction of Log Mel Spectrogram features, the window size and hop size of the STFT are set to 0.025 and 0.010 s, respectively. The 96 × 64 Log Mel Spectrogram is then input into VGGish. The data augmentation strategy for audio primarily involves random cropping along the audio time axis. The output remains consistent with that of the face recognition task, yielding a 128-dimensional representation for speech.

During the pretraining process, this work employs a shared-parameter classifier to jointly train the two modalities. We define the training data as $D = \{(f_i, v_i, y_i)\}_{i=1}^{N}$, where f_i represents the facial image embedding generated by the face recognition network belonging to identity y_i, and v_i denotes the voice segment embedding of identity y_i. The weight matrix of the classifier is represented as $W = [\omega_1, \omega_2, \cdots, \omega_N] \in R^{D\times N}$, where N and D represent the number of identities and feature dimensions, respectively. Given the strong performance of norm-softmax in classification tasks [20], the weight matrix is normalized. The learning objective for the pre-training phase is defined as follows:

$$L_{ID} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\omega_{y_i}^{T} f_i\right)}{\sum_{j=1}^{M}\exp\left(\omega_j^{T} f_i\right)} - \frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\omega_{y_i}^{T} v_i\right)}{\sum_{j=1}^{M}\exp\left(\omega_j^{T} v_i\right)} s.t. \left\|w_j\right\| = r \tag{1}$$

where f_i and v_i represent the face features and speech features corresponding to identity y_i, respectively. The losses from both modalities are calculated and jointly backpropagated to update the network parameters. Track the loss L_{ID} during the training process to ensure it gradually decreases and stabilizes.

This method integrates identity supervision into joint training to achieve unified representations, providing an effective solution for cross-modal tasks. Furthermore, it establishes a scalable foundation for other applications, such as emotion analysis and multimodal recognition. Subsequently, the pre-trained network is utilized to encode facial images and speech segments, facilitating the development of a cross-modal alignment module for these features.

3.2 Cross-Modal Simplex Center Loss

Most existing methods primarily align face and speech features by optimizing the distance between sample pairs. In contrast, we introduce a cross-modal center loss to model the many-to-many relationship between speech and face, leveraging global information to learn and minimize the distance between samples and their corresponding class centers.

Given N classes and M modalities of feature embeddings $\{e_i^m\}_{i=1}^{N}$ (where $m \in [1, M]$), the single-modal center loss [36] is directly extended, leading to the formulation of the cross-modal center loss [21] as shown in Eq. (2):

$$L_c = \frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{M}\left\|e_i^m - C_{y_i}\right\|_2^2 \tag{2}$$

where $C_{y_i} \in R^D$ represents the center of identity y_i in the semantic space, where D is the feature dimension. However, this cross-modal loss still relies on the softmax loss function and its variants to learn the spacing between different classes. Furthermore, the data distributions across different modalities exhibit significant differences, and aligning centers without cross-modal interaction inevitably hampers the generalization performance of the model.

To this end, we introduce a simplex classifier [37–39] to enhance the cross-modal center loss. Studies have shown that high-dimensional data tends to distribute near the outer surface of a hypersphere, and as the

dimensionality increases, the data points converge toward the vertices of a regular simplex inscribed within the hypersphere [40]. We leverage the vertices of a regular simplex inscribed in the hypersphere as identity centers, driving the alignment of face and speech embeddings around their corresponding identity centers. When the feature dimension D is greater than or equal to the number of categories minus one ($D \geq N-1$), a regular N-simplex can be constructed circumscribed about the hypersphere. Assuming the radius of the hypersphere is 1, the vertices of the regular simplex are defined as:

$$C_i = \begin{cases} (N-1)^{-\frac{1}{2}}\mathbf{1}, i = 1 \\ \kappa\mathbf{1} + \eta e_{i-1}, 2 \leq i \leq N \end{cases} \tag{3}$$

where

$$\kappa = -\frac{1+\sqrt{N}}{(N-1^{\frac{3}{2}})}, \eta = \sqrt{\frac{N}{N-1}} \tag{4}$$

Here, $\mathbf{1}$ denotes a vector of all ones, and e_i represents the natural basis vector with a 1 in the i-th position and 0 s elsewhere.

The radius r of the hypersphere can serve as a scaling parameter in the optimization objective and should be set to an appropriate value greater than 1 in high-dimensional spaces. Specifically, for the speech-face association task, the identity centers to be learned are defined as follows:

$$C_i = rx_i, i = 1, \ldots, N \tag{5}$$

Here, x_i represents the vertices of a regular N-simplex circumscribed about a unit hypersphere. According to the properties of a regular simplex, the distances between vertices are equal and remain invariant under rotation or translation. By learning to align embeddings from different modalities with their respective identity centers, the model simultaneously optimizes inter-class distances in both Euclidean space and angular space.

Considering the distribution differences between face and speech modal embeddings in the latent space, we employ a cross-modal projection mechanism during the alignment process of the simplex centers. This mechanism projects the feature vector of one modality onto the direction of the feature vector of the other modality. Subsequently, using the vertices of the regular N-simplex as identity centers, we perform center alignment on these cross-modal projection vectors. The final model optimizes the many-to-many relationship between speech and face features in the common identity space using Eq. (6):

$$L_c = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{f}_i - C_{y_i}\right\|_2^2 + \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{v}_i - C_{y_i}\right\|_2^2 \tag{6}$$

where $\hat{f}_i$ represents the vector projection of the face feature f_i onto the normalized speech feature $\overline{v}_i$, while $\hat{v}_i$ denotes the vector projection of the speech feature v_i onto the normalized face feature $\overline{f}_i$. As illustrated in Fig. 2b, opposite directions of the two vectors will result in negative scalar projections, which justifies the multiplication by the two normalized features in the formula.

$$\hat{f}_i = {f_i}^T\overline{v}_i \cdot \overline{v}_i, \hat{v}_i = {v_i}^T\overline{f}_i \cdot \overline{f}_i \tag{7}$$

Here, $\overline{v} = \frac{v}{\|v\|}$ and $\overline{f} = \frac{f}{\|f\|}$. As the direction of the face and speech feature vectors becomes more similar, the scalar of the projection vector increases. This process is consistent with the alignment of simplex centers, optimizing both Euclidean space and angular space distances simultaneously. This cross-modal projection

operation, which focuses on speech and face embeddings, aims to enhance the correlation between different modalities. It mitigates the significant differences in direction and amplitude between the two modalities in the shared space, thereby accelerating the central alignment process. By optimizing the distance between the cross-modal projections and the centers, it ensures that features from different modalities are effectively aligned, improving retrieval accuracy.

Compared to the original cross-modal center loss, the proposed cross-modal simplex center loss provides more robust centers and takes into account the interactions between modalities, enhancing intra-class compactness while maximizing inter-class distances.

3.3 Triplet Center Loss

In the task of speech-face association, it is essential not only to align features based on identity categories but also to learn discriminative features between speech and facial samples from different identities. Intuitively, large-scale face and speech datasets often contain hard negative samples, such as individuals with similar facial appearances but distinct voices, and *vice versa*. To handle these challenging cases, inspired by the triplet loss [41,42], we construct triplets consisting of a query, a positive sample, and a hard negative sample during model training. This approach enhances the model's ability to distinguish between samples of visually or audibly similar individuals.

In this framework, the identity center C_{y_i} mentioned in Section 3.2 is set as the query. During model training, each mini-batch contains N identities, with each identity represented by a pair of positive speech-face samples, while samples from different identities act as negatives for one another. The local relationships between sample pairs are optimized using the following loss function:

$$L_{dis} = \sum_{i=1}^{N} \left[m + < f_i, C_{y_i} > - \min_{j \neq y_i} < f_j, C_{y_i} > \right]_+ + \sum_{i=1}^{N} \left[m + < v_i, C_{y_i} > - \min_{j \neq y_i} < v_j, C_{y_i} > \right]_+ \tag{8}$$

Here, m is the margin, $< \cdot, \cdot >$ denotes the distance metric, which is implemented as Euclidean distance in this work, and $[x]_+ = \max(0, x)$. Since the number of identities in the dataset far exceeds the batch size, each query sample has only a few positive examples within a batch. To address the imbalance between positive and negative samples and mitigate its adverse impact on gradient optimization, the model penalizes only the hardest negative sample, i.e., the one with the highest similarity to the query identity center.

Under the training of the triplet-center loss, samples from the same category demonstrate high similarity, while samples from different categories exhibit low similarity. Through the joint training of the cross-modal center loss and triplet center loss, we effectively learn modality-independent and discriminative features.

$$Loss = L_c + \alpha L_{dis} \tag{9}$$

In the formula, α is a hyperparameter for the weight of the triplet center loss.

3.4 Optimization

In this subsection, we provide a detailed and step-by-step explanation of the optimization process for the proposed speech-face cross-modal association learning framework. The optimization details are outlined in Algorithm 1. The entire process is divided into two stages: joint pretraining and cross-modal simplex center alignment.

Algorithm 1: The proposed method

Input:

Dataset D: face image and voice pairs $\{x_i^f, x_i^v, y_i\}_{i=1}^N$;

Simplex center parameters: radius of the hypersphere r, total number of identities N;

Training hyperparameters: learning rate lr, margin m, loss weight α;

Step 1: Joint Pre-Training

1. Randomly initialize the face network θ_f, voice network θ_v, and classifier weight W;
2. **While** not converged **do**:

   Extracting face embeddings $f_i = f_{\theta_f}\left(x_i^f\right)$;

   Extracting voice embeddings $v_i = f_{\theta_v}(x_i^v)$;

   Compute Identity Supervision Loss L_{ID} by Eq. (1);

   Update θ_f, θ_v, W by descending the gradient;

   End While;

Step 2: Cross-Modal Simplex Center Alignment

1. Initialize the face encoder θ_f, voice encoder θ_v with the joint pre-training results;
2. Create regular simplex by Eqs. (3)–(5);
3. Extracting embeddings $f_i' = f_{\theta_f}\left(x_i^f\right)$, $v_i' = f_{\theta_v}(x_i^v)$;
4. While not converged do:

   Re-encoding $f_i = f_{\varphi_f}(f_i')$, $v_i = f_{\varphi_v}(v_i')$;

   Compute Cross-Modal Simplex Center Loss L_C by Eq. (6);

   Compute Cross-Modal Triplet Center Loss L_{dis} by Eq. (8);

   Compute total loss $Loss$ by Eq. (9);

   Update φ_f, φ_v by descending the gradient;

   End While;

Return: Optimized face encoder φ_f, voice encoder φ_v;

First, the shared parameters W of the classifier are randomly initialized. VGGNet and VGGish are jointly pretrained on the training set under identity supervision, resulting in the optimized network parameters θ_f and θ_v after iterative updates. Next, the pretrained parameters are used to initialize and freeze the face and speech networks, which are then employed to extract face and speech embeddings as inputs for the cross-modal simplex center alignment module. Subsequently, the extracted embeddings undergo further encoding and alignment, utilizing face and speech encoders with parameters φ_f and φ_v, respectively. The optimization of these encoder parameters is guided by a weighted combination of the cross-modal simplex center loss and triplet center loss, as defined in Eq. (9).

4 Experiment

4.1 Datasets

Following prior works [31,32,43], we evaluated the proposed method using publicly available datasets, VoxCeleb [44] and VGGFace [34], which were originally introduced in these references. VoxCeleb is a large-scale human speech dataset collected from YouTube, containing speech-video pairs from 1251 identities, while VGGFace is a face dataset consisting of facial images from 2622 identities.

To conduct our experiments, we utilized the overlapping subset of these datasets as described in the referenced works, which includes 1225 shared identities with high-quality speech and facial image data. The face images and speech segments were partitioned into training, validation, and test sets based on the

protocol specified in these references. The partitions were carefully designed to ensure that identities in the training, validation, and test sets do not overlap, thereby preventing any identity leakage during evaluation.

The division of the dataset is summarized in Table 1. Specifically, the training set was used for model learning, the validation set for hyperparameter tuning and performance monitoring, and the test set for final evaluation. Queries for validation and testing were generated following the evaluation protocol outlined in Section 4.3.

Table 1: Details relating to the datasets

	Facial images	Voice segments	Identities
Train	104,724	113,322	924
Validation	12,260	14,182	112
Test	20,076	21,850	189
Total	137,060	149,354	1225

4.2 Implementation Details

The training process consists of two stages: a pre-training stage for joint face recognition and speaker recognition, followed by a feature alignment stage where the recognition network is frozen. The implementation details of the pre-training network are described in Section 3.1. For the raw data, images are cropped to $112 \times 112 \times 3$ to capture the full face, while voice segments are resampled at 16,000 Hz and randomly cropped to a duration between 2.5 to 5.0 s, then converted into log-Mel spectrograms. After pre-training, the image data and speech data are encoded into 128-dimensional face features and speaker features, respectively. In the feature alignment stage, both the structure of the vocal system and language habits can reflect a person's identity, so the face and speech encoders are designed similarly, consisting of two stacked self-attention blocks. This design facilitates the learning of modality-independent identity information, which is then concatenated with a multi-layer perceptron (MLP) layer to expand the dimensionality to 1024, satisfying the condition $D \geq N - 1$.

To ensure the model's generalization ability in the association task, we randomly sample multiple identities during each iteration to maintain balanced participation in training. For each selected identity, we then randomly choose one face image and its corresponding speech audio segment as the training samples.

The proposed method is implemented using PyTorch. For model training, we utilize a Stochastic Gradient Descent (SGD) optimizer with a batch size of 64 and a momentum of 0.9. The learning rate is initialized at 0.1 and decays by a factor of 0.1 at the 2000 and 3000 iteration marks. Regarding parameter settings, we set $r = 90$ (Eq. (4)), $m = 3$ (Eq. (7)), and $\alpha = 0.2$ (Eq. (8)), with a maximum iteration limit of 10,000. The best-performing model on the validation set is retained for evaluation.

4.3 Testing Protocol

We evaluate the overall effectiveness of the proposed cross-modal center learning framework in the speech-face association learning task through three key tasks. For each task, we assess the model's performance in both voice-to-face (V-F) and face-to-voice (F-V) directions. In the matching and verification tasks, we incorporate gender constraint settings to evaluate the model's performance in more challenging scenarios.

1:N Matching. In this task, a voice segment or a face image is used as a query, and the goal is to identify the corresponding instance from N candidate instances in the other modality, with only one candidate matching the query identity. In the experiments, the value of N varies from 2 to 10, and cosine similarity is employed for comparison to assess the test accuracy (ACC) across different values of N.

Verification. In this task, a face image and a voice segment are presented as inputs, and the objective is to determine whether they correspond to the same identity, effectively framing it as a binary classification problem. The performance is evaluated using the Area Under the ROC Curve (AUC).

Retrieval. In this task, a face image or a voice segment serves as the query, and candidates from the respective collections are ranked. The gallery contains one or more instances that correspond to the query. The goal is to rank the candidates such that those matching the query appear at the top. Performance is evaluated using Mean Average Precision (mAP).

4.4 Results and Comparison

Comparison Experiments. To validate the superiority of the proposed method, we compared it with well-established baseline approaches and subsequent improved models, including SVHF [6], DIMNet [32], Wang's model [43], and Wen's model [31]. These methods have demonstrated strong performance in prior works and are widely recognized within the research community. By benchmarking against these competitive approaches, we aimed to provide a comprehensive evaluation of our method's effectiveness in addressing the challenges of cross-modal association and retrieval. It should be noted that SVHF-Net achieves forced matching through a three-stream convolutional neural network and feature fusion. However, this architecture is not suitable for cross-modal verification and retrieval tasks. DIMNet leverages common covariates between speech and face, using a modality-shared classifier for non-joint mapping. Both Wang's and Wen's models adopt a joint learning strategy combining identity loss and pairwise loss. Wang's model introduces hard negative sample mining, while Wen's model incorporates an adaptive identity reweighting mechanism. Table 2 highlights the performance of our proposed simplex center learning method relative to these existing approaches in 1:2 matching and verification tasks. The best results are shown in bold. In this context, 'G' denotes gender-constrained scenarios, where both the query and candidate samples belong to the same gender, whereas 'U' indicates unrestricted genders.

Table 2: Comparison of model performance in 1:2 matching, verification and retrieval

Methods	1:2 Matching (ACC)				Verification (AUC)				Retrieval (mAP)	
	V-F(U)	F-V(U)	V-F(G)	F-V(G)	V-F(U)	F-V(U)	V-F(G)	F-V(G)	V-F	F-V
SVHF [6]	80.15	78.96	63.29	63.01	–	–	–	–	–	–
DIMNet [32]	81.02	81.46	69.65	69.33	80.49	81.21	69.51	68.76	4.08	3.64
Wang's [43]	83.51	84.23	72.07	71.58	82.54	82.80	70.26	70.09	4.35	3.32
Wen's [31]	86.92	86.15	76.74	75.21	86.92	86.64	76.47	75.79	5.32	5.51
Ours	**88.23**	**87.74**	**79.22**	**78.43**	**88.19**	**88.07**	**78.29**	**78.11**	**5.87**	**5.95**

As shown in Table 2, our method outperforms the compared methods in 1:2 matching, verification, and retrieval tasks. Specifically, without gender constraints, the proposed method improves by 1.3%–8.8% in matching and verification tasks, and by 0.6%–2.5% in retrieval, demonstrating stable performance whether from voice to face or from face to voice. In more challenging gender-constrained scenarios, the model demonstrates a significant improvement in accuracy. Particularly in the face-to-speech matching task, it achieves a 3.4%–7.0% improvement compared to the recent works of Wang and Wen, outperforming the baseline methods by 15.6%. Our analysis suggests that although Wang's approach employs pairwise hardest negative mining for discriminative feature learning, our improved triplet center loss directly uses identity centers as queries to mine the hardest negatives, resulting in a more stable learning process. In contrast, Wen's adaptive identity reweighting mechanism inevitably leads to information loss, causing a noticeable drop in

performance when handling challenging samples. Fig. 3 displays examples of constrained matching results, indicating that our method effectively learns various identity-related deep features beyond gender from faces and voices, revealing profound associations between them.

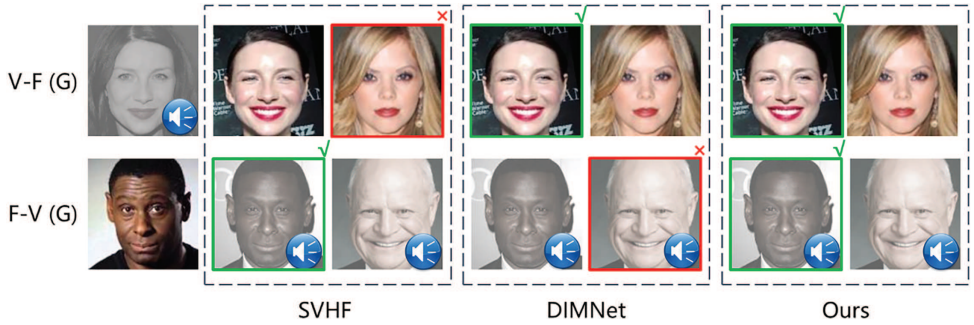

Figure 3: Qualitative experimental results under gender constraints

As illustrated in Fig. 4, the 1:N matching task becomes increasingly challenging as N grows, resulting in a general decline in accuracy across all methods. However, our proposed method (red curve) consistently demonstrates higher accuracy throughout the process, with a relatively minor decrease. This suggests that our approach exhibits greater robustness and adaptability in larger-scale matching tasks. This advantage is primarily attributed to the proposed cross-modal simplex center learning framework, which effectively captures the many-to-many relationships between faces and voices, thereby sustaining superior matching performance in a broader search space. Moreover, in gender-constrained matching tasks, the performance of all models tends to be lower than in scenarios without gender constraints. We attribute this to the presence of more similar samples from different identities within the same gender, making the task inherently more difficult. Nevertheless, our method still outperforms the others in these challenging conditions, reinforcing its potential advantages in addressing complex matching tasks.

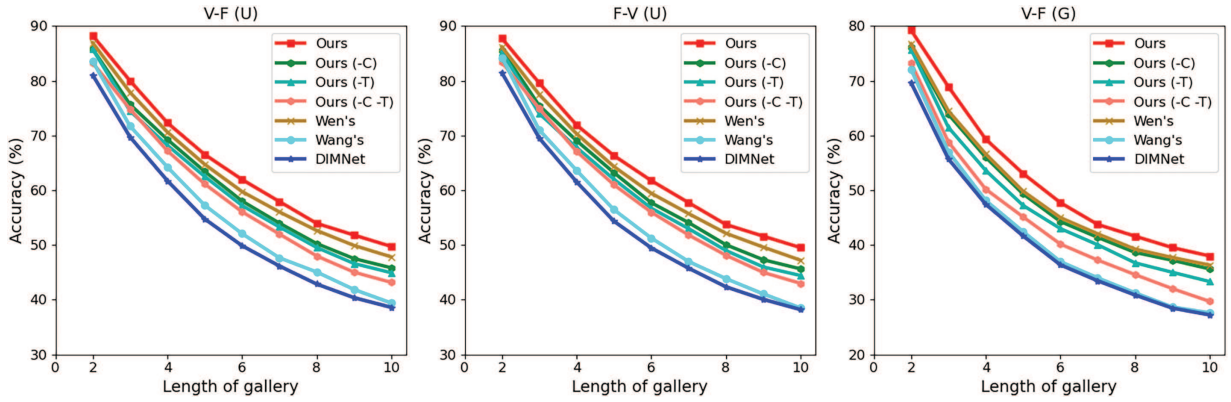

Figure 4: Quantitative results on 1:N matching task

Ablation Experiments. To assess the effectiveness of the proposed speech-face association learning framework, we conducted ablation studies focusing on two key components: the cross-modal simplex center loss and the triplet center loss. This led to three ablated variants: (1) a model utilizing the original Cross-modal Center Loss (CCL) instead of our proposed cross-modal simplex center loss (denoted as -C in Fig. 3),

(2) a model with the Triplet Center Loss (TCL) omitted (denoted as -T in Fig. 3), and (3) a model using the original CCL while removing the triplet center loss, resulting in a model (denoted as -C -T in Fig. 3). This configuration allows us to evaluate the impact of both losses on the model's overall performance. The results are presented in Table 3, highlighting the accuracy of the model and its variants in the 1:2 matching task for clearer comparison. The best results are shown in bold.

Table 3: Ablation experiment results of 1:2 matching

Methods	V-F(U)	F-V(U)	V-F(G)	F-V(G)
w/original CCL	85.79	85.63	76.14	75.87
w/o TCL	85.76	85.27	75.74	75.63
w/o CSCL&TCL	83.26	83.49	73.26	73.01
Ours	**88.23**	**87.74**	**79.22**	**78.43**

As shown in Fig. 4, when not combined with triplet center loss, the proposed cross-modal simplex center loss outperforms the original cross-modal center loss in 1:N matching tasks, and this advantage becomes more pronounced as N increases. We speculate that, compared to the original method, the proposed method provides stronger robustness by maximizing the distance between identity centers. After combining with triplet center loss, this performance advantage is further expanded, confirming that our method achieves effective cross-modal interaction. The learned face and voice embeddings cluster closely around the identity centers, demonstrating a clear advantage in inter-class distinction.

In Table 3, we observe that even with only the joint pretraining model, it maintains higher accuracy and stability compared to baseline methods, reaffirming that parameter-sharing unimodal classification aids cross-modal association. Notably, in the gender-constrained 1:2 matching task, the triplet center loss provides substantial performance enhancements, indicating that hard negative mining effectively captures deep associations between faces and voices.

However, attempts to increase the weight of the triplet center loss during experimentation resulted in significant performance degradation. We speculate that this may cause the model to focus excessively on hard negative samples, leading to overfitting. Therefore, enhancing the model's generalization ability and robustness when dealing with complex samples remains an area for further exploration.

5 Conclusion

In this paper, we present a novel speech-face association framework based on cross-modal simplex center learning, which effectively enhances feature alignment between speech and facial images. By employing a joint pretraining approach, our method leverages well-established face recognition and speaker recognition techniques, significantly reducing the model's training costs while ensuring robust performance. The combination of cross-modal simplex center loss and improved triplet center loss, has proven effective in enhancing both intra-class compactness and inter-class separability of cross-modal features. Quantitative evaluations validate the superiority of our method, compared with existing methods, our method achieved improvements in matching, validation, and retrieval tasks, especially in challenging gender constrained experiments, with an accuracy increase of 2.48%–15.93%. These results underline the practical potential of our framework for real-world applications in cross-modal feature alignment.

Future work will focus on refining the framework to address inter-modal discrepancies further, such as by integrating advanced self-supervised learning or domain adaptation techniques. Additionally, we aim to explore its applicability to other challenging cross-modal scenarios, including emotion recognition and

multi-modal biometric authentication, thereby extending the impact of this research to a broader range of applications.

Acknowledgment: The authors would like to express their gratitude and appreciation to People's Public Security University of China for all the support and facilities throughout the research.

Funding Statement: This project was partially funded by the Scientific Funding for China Academy of Railway Sciences Corporation Limited, China (No. 2023YJ125).

Author Contributions: Conceptualisation, Qiming Ma; methodology, Qiming Ma and Fanliang Bu; software, Qiming Ma; validation, Qiming Ma and Lingbin Bu; formal analysis, Qiming Ma and Yifan Wang; investigation, Qiming Ma; resources, Qiming Ma and Zhiyuan Li; writing—original draft preparation, Qiming Ma; writing—review and editing, Qiming Ma and Rong Wang; supervision, Fanliang Bu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Kärkkäinen K, Joo J. FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV); 2021. p. 1547–57. doi:10.1109/WACV48630.2021.00159.
2. Hechmi K, Trong TN, Hautamäki V, Kinnunen T. Voxceleb enrichment for age and gender recognition. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU); 2021. p. 687–93. doi:10.48550/arXiv.2109.13510.
3. Kamachi M, Hill H, Lander K, Vatikiotis-Bateson E. Putting the face to the voice: matching identity across modality. Curr Biol. 2003;13(19):1709–14. doi:10.1016/j.cub.2003.09.005.
4. Rhone AE, Rupp K, Hect JL, Harford E, Tranel D, Howard MA, et al. Electrocorticography reveals the dynamics of famous voice responses in human fusiform gyrus. J Neurophysiol. 2022 Dec;129(2):342–6. doi:10.1152/jn.00459.2022.
5. Nagrani A, Albanie S, Zisserman A. Learnable PINs: cross-modal embeddings for person identity. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018; Munich, Germany. p. 73–89. doi:10.48550/arXiv.1805.00833.
6. Nagrani A, Albanie S, Zisserman A. Seeing voices and hearing faces: cross-modal biometric matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 8427–36. doi::10.1109/CVPR.2018.00879.
7. Horiguchi S, Kanda N, Nagamatsu K. Face-voice matching using cross-modal embeddings. In: Proceedings of the 26th ACM International Conference on Multimedia; Seoul, Republic of Korea; 2018. p. 1011–9. doi:10.1145/3240508.3240601.
8. Kim C, Shin HV, Oh T-H, Kaspar A, Elgharib M, Matusik W. On learning associations of faces and voices. In: Computer Vision-ACCV: 14th Asian Conference on Computer Vision; 2019; Perth, Australia. p. 276–92. doi:10.1007/978-3-030-20873-8_18.
9. He X, Zhou Y, Zhou Z, Bai S, Bai X. Triplet-center loss for multi-view 3D object retrieval. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018; Salt Lake City, UT, USA. p. 1945–54. doi:10.1109/CVPR.2018.00208.

10. Yang S, Tantrawenith M, Zhuang H, Zhiyong W, Aolan S, Jianzong W, et al. Speech representation disentanglement with adversarial mutual information learning for one-shot voice conversion. doi:10.48550/arXiv.2208.08757.
11. Qu L, Liu M, Wang W, Zheng Z, Nie L, Chua T-S. Learnable pillar-based re-ranking for image-text retrieval. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2023; New York, NY, USA. p. 1252–61. doi:10.1145/3539618.3591712.
12. Fan Y, Lin Z, Saito J, Wang W, Komura T. FaceFormer: speech-driven 3D facial animation with transformers. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022; New Orleans, LA, USA. p. 18749–58. doi:10.1109/CVPR52688.2022.01821.
13. Cao M, Li S, Li J, Nie L, Zhang M. Image-text retrieval: a survey on recent research and development. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence; 2022; Vienna, Austria. p. 5410–7. doi:10.24963/ijcai.2022/755.
14. Wang X, Li L, Li Z, Wang X, Zhu X, Wang C, et al. Aligning cross-modal entities for image-text retrieval upon vision-language pre-trained models. In: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining; 2023; New York, NY, USA. p. 456–64. doi:10.1145/3539597.3570481.
15. Xia Y, Huang H, Zhu J, Zhao Z. Achieving cross modal generalization with multimodal unified representation. In: Advances in Neural Information Processing Systems; 2023; New Orleans, LA, USA. p. 63529–41.
16. Kim D, Saito K, Oh T-H, Plummer BA, Sclaroff S, Saenko K. CDS: cross-domain self-supervised pre-training. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021. p. 9103–12. doi:10.1109/ICCV48922.2021.00899.
17. Fu Z, Mao Z, Song Y, Zhang Y. Learning semantic relationship among instances for image-text matching. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023; Los Angeles, CA, USA. p. 15159–68. doi:10.1109/CVPR52729.2023.01455.
18. You H, Zhou L, Xiao B, Codella N, Cheng Y, Xu R, et al. Learning visual representation from modality-shared contrastive language-image pre-training. In: Computer Vision-ECCV 2022: 17th European Conference; 2022; Tel Aviv, Israel. p. 69–87. doi:10.1007/978-3-031-19812-0_5.
19. Jiang D, Ye M. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023; Los Angeles, CA, USA. p. 2787–97. doi:10.1109/CVPR52729.2023.00273.
20. Zhang Y, Lu H. Deep cross-modal projection learning for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018; Munich, Germany. p. 707–23. doi:10.1007/978-3-030-01246-5_42.
21. Peng Y, Qi J. CM-GANs: cross-modal generative adversarial networks for common representation learning. ACM Trans Multimedia Comput Commun Appl. 2019;15(1):22. doi:10.1145/3284750.
22. Sun C, Chen M, Cheng J, Liang H, Zhu C, Chen J, et al. Supervised cross-modal contrastive learning for audio-visual coding. In: Proceedings of the 31st ACM International Conference on Multimedia; 2023; New York, NY, USA. p. 261–70. doi:10.1145/3581783.3613805.
23. Yan J, Deng C, Huang H, Liu W. Causality-invariant interactive mining for cross-modal similarity learning. IEEE Trans Pattern Anal Mach Intell. 2024;46(9):6216–30. doi:10.1109/TPAMI.2024.3379752.
24. Liu AH, Jin S, Lai C-I, Rouditchenko A, Oliva A, Glass JR. Cross-modal discrete representation learning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics; 2021; Dublin, Ireland. p. 3013–35. doi:10.18653/v1/2022.acl-long.215.
25. Jing L, Vahdani E, Tan J, Tian Y. Cross-modal center loss for 3D cross-modal retrieval. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021. p. 3141–50. doi:10.1109/CVPR46437.2021.00316.
26. Surís D, Duarte A, Salvador A, Torres J, Giró-i-Nieto X. Cross-modal embeddings for video and audio retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops; 2018; Berlin, Germany. p. 711–6. doi:10.1007/978-3-030-11018-5_62.
27. Gabeur V, Nagrani A, Sun C, Alahari K, Schmid C. Masking modalities for cross-modal video retrieval. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2022; Santa Monica, CA, USA. p. 2111–20. doi:10.1109/WACV51458.2022.00217.

28. Hao X, Zhang W, Wu D, Zhu F, Li B. Listen and look: multi-modal aggregation and co-attention network for video-audio retrieval. In: 2022 IEEE International Conference on Multimedia and Expo (ICME); 2022; Yokohama, Japan. p. 1–6. doi:10.1109/ICME52920.2022.9859647.
29. Yuan Z, Shen Q, Zheng B, Liu Y, Jiang L, Guo G. Video and audio are images: a cross-modal mixer for original data on video-audio retrieval. Knowl Based Syst. 2024;299:112076. doi:10.1016/j.knosys.2024.112076.
30. Nawaz S, Janjua MK, Gallo I, Mahmood A, Calefati A. Deep latent space learning for cross-modal mapping of audio and visual signals. In: 2019 Digital Image Computing: Techniques and Applications (DICTA); 2019; Brisbane, Australia. p. 1–7. doi:10.1109/DICTA47822.2019.8945863.
31. Wen P, Xu Q, Jiang Y, Yang Z, He Y, Huang Q. Seeking the shape of sound: an adaptive framework for learning voice-face association. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 16347–56. doi:10.1109/DICTA47822.2019.8945863.
32. Wen Y, Al Ismail M, Liu W, Raj B, Singh R. Disjoint mapping network for cross-modal matching of voices and faces. In: ICLR 2019; 2019 [cited 2024 Nov 11]. p. 1–17. Available from: https://api.semanticscholar.org/Corpus.
33. Zheng A, Hu M, Jiang B, Huang Y, Yan Y, Luo B. Adversarial-metric learning for audio-visual cross-modal matching. IEEE Trans Multimed. 2022;24(9):338–51. doi:10.1109/TMM.2021.3050089.
34. Parkhi OM, Vedaldi A, Zisserman A. Deep face recognition. In: British Machine Vision Conference; 2015 [cited 2024 Nov 11]; Cambridge, UK. Available from: https://api.semanticscholar.org/CorpusID:4637184.
35. Hershey S, Chaudhuri S, Ellis DPW, Gemmeke JF, Jansen A, Moore RC, et al. CNN architectures for large-scale audio classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2017; New Orleans, LA, USA. p. 131–5. doi:10.1109/ICASSP.2017.7952132.
36. Wen Y, Zhang K, Li Z, Qiao Y. A discriminative feature learning approach for deep face recognition. In: Computer Vision-ECCV 2016: 14th European Conference; 2016; Amsterdam, The Netherlands. p. 499–515. doi:10.1007/978-3-319-46478-7_31.
37. Pernici F, Bruni M, Baecchi C, Bimbo AD. Regular polytope networks. IEEE Trans Neural Net Learn Syst. 2022;33(9):4373–87. doi:10.1109/TNNLS.2021.3056762.
38. Bytyqi Q, Wolpert N, Schömer E, Schwanecke U. Prototype softmax cross entropy: a new perspective on softmax cross entropy. In: Image analysis. Cham: Springer; 2023. p. 16–31. doi:10.1007/978-3-031-31438-4_2.
39. Cevikalp H, Saribas H. Deep simplex classifier for maximizing the margin in both Euclidean and angular spaces. In: Image analysis. Cham: Springer; 2023. p. 91–107. doi:10.1007/978-3-031-31438-4_7.
40. Hall P, Marron JS, Neeman A. Geometric representation of high dimension, low sample size data. J R Stat Soc Ser B: Stat Methodol. 2005;67(3):427–44. doi:10.1111/j.1467-9868.2005.00510.x.
41. Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015; Boston, MA, USA. p. 815–23. doi:10.1109/CVPR.2015.7298682.
42. Khan M, Saeed M, Saddik AEl, Gueaieb W. ARTriViT: automatic face recognition system using ViT-based siamese neural networks with a triplet loss. In: 2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE); 2023; Helsinki, Finland. p. 1–6. doi:10.1109/ISIE51358.2023.10228106.
43. Wang R, Liu X, Cheung Y, Cheng K, Wang N, Fan W. Learning discriminative joint embeddings for efficient face and voice association. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval; 2020. p. 1881–4. doi:10.1145/3397271.3401302.
44. Nagrani A, Chung JS, Zisserman A. VoxCeleb: a Large-Scale Speaker Identification Dataset. 2017. doi:10.48550/arXiv.1706.08612.