Flexible Biometrics Recognition: Bridging the Multimodality Gap through Attention, Alignment and Prompt Tuning

Leslie Ching Ow Tiong[*,1] Dick Sigmund[*,2] Chen-Hui Chan[3] Andrew Beng Jin Teoh[†,4]
[1]Samsung Electronics [2]AIDOT Inc. [3]Korea Institute of Science and Technology [4]Yonsei University
[1]leslie.tiong@samsung.com [2]dsigmund@aidot.ai [3]chchan@kist.re.kr [4]bjteoh@yonsei.ac.kr

Abstract

Periocular and face are complementary biometrics for identity management, albeit with inherent limitations, notably in scenarios involving occlusion due to sunglasses or masks. In response to these challenges, we introduce Flexible Biometric Recognition (FBR), a novel framework designed to advance conventional face, periocular, and multimodal face-periocular biometrics across both intra- and cross-modality recognition tasks. FBR strategically utilizes the Multimodal Fusion Attention (MFA) and Multimodal Prompt Tuning (MPT) mechanisms within the Vision Transformer architecture. MFA facilitates the fusion of modalities, ensuring cohesive alignment between facial and periocular embeddings while incorporating soft-biometrics to enhance the model's ability to discriminate between individuals. The fusion of three modalities is pivotal in exploring interrelationships between different modalities. Additionally, MPT serves as a unifying bridge, intertwining inputs and promoting cross-modality interactions while preserving their distinctive characteristics. The collaborative synergy of MFA and MPT enhances the shared features of the face and periocular, with a specific emphasis on the ocular region, yielding exceptional performance in both intra- and cross-modality recognition tasks. Rigorous experimentation across four benchmark datasets validates the noteworthy performance of the FBR model. The source code is available at https://github.com/MIS-DevWorks/FBR.

1. Introduction

Facial recognition has attained ubiquitous applications across diverse domains today [18, 19, 35]. However, they struggle with challenges arising from cosmetic changes, plastic surgery, and particularly obstructions like face masks. On the other hand, periocular recognition, which focuses on the region around the eyes, has gained traction as an alternative to face recognition [1, 23–25]. Despite the significant progress, however, challenges remain, especially with glasses or sunglasses, impacting the accuracy of periocular recognition.

The fusion of facial and periocular [15, 32, 37], holds promise for enhancing recognition performance. However, traditional multimodal biometrics present fresh challenges in managing and storing templates of all biometric modalities, which can result in computational and storage overhead. Moreover, ensuring all modalities are available for recognition is critical for seamless deployment. In response to these challenges, conditional biometrics [11, 22], along with cross-modality biometrics recognition [20, 33], offer promising avenues to mitigate the constraints of unimodal biometric systems, i.e., sole face or periocular recognition as well as multimodal biometric systems. Conditional biometrics enhance a single modality by incorporating information from another, such as periocular recognition conditioned by the face or vice versa. On the other hand, cross-modality biometrics encompasses the task of matching biometric samples across distinct biometric modalities, such as face vs. periocular matching.

This paper introduces Flexible Biometrics Recognition (FBR), designed to support intra- and cross-modality biometric matching, as illustrated in Figure 1. The FBR model is initially trained to align facial, periocular, and soft-biometric attributes. The latter encompasses social or physical descriptive traits of individuals such as gender, age, or ethnicity, which have proven to enhance the discriminative power of the embedding [6]. During deployment, the trained FBR model serves as a feature extractor, acquiring facial or periocular embeddings based on the input modality. In contrast to unimodal and multimodal biometric systems, FBR produces modality-invariant embeddings, facilitating both intra- and cross-modality matching. Additionally, FBR can address scenarios where only facial or periocular data is stored, enabling exclusive reliance on the available modality during recognition.

The adaptability of FBR ensures robust performance and operational reliability, even in incomplete or temporarily

[*]The authors have contributed equally to this work.
[†]Corresponding author.

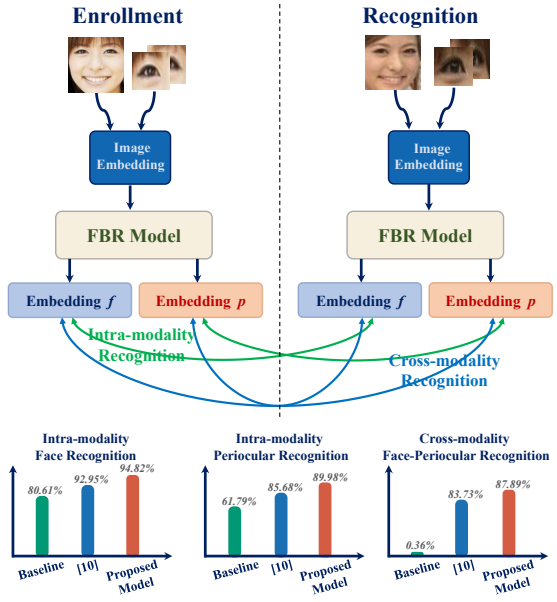

Figure 1. An overview of the Flexible Biometrics Recognition (FBR). This approach encompasses flexible recognition tasks such as intra- and cross-modality biometric recognition, demonstrating its versatility in biometric systems. Our model exhibits enhanced performance relative to benchmarks, including unimodal biometrics baseline and a competing model [10].

unavailable biometric data. As depicted in Figure 1, the proposed FBR model outperforms the unimodal biometrics baseline and a competing model [10], which is also trained with three modalities and equips with a prompt tuning mechanism. While FBR offers exceptional adaptability, attaining decent performance in intra- and cross-modality matching is a non-trivial challenge. Enhancing one facet could potentially undermine the other and vice versa. Striking the right balance between optimizing intra- and cross-modality recognition is crucial.

To substantiate FBR, we introduce a Multimodal Fusion Attention Vision Transformer (MFA-ViT) and a multimodal-prompt tuning (MPT) mechanism. MFA-ViT is crafted to establish a cohesive alignment between facial and periocular features within a ViT. Furthermore, integrating soft-biometric attributes enhances the model's ability to discern differences between identities effectively. The multimodal fusion attention (MFA) module achieves the fusion of three modalities, which is pivotal in exploring interrelationships between different modalities. This endeavor proves advantageous in mitigating the trade-off between intra- and cross-modality recognition tasks.

In the realm of robust embedding learning across diverse modalities, prompt tuning has proven effective for modality alignment [10, 39]. Nevertheless, standard prompt tuning (SPT) often falls short of effectively utilizing multimodal information. For FBR problems, SPT lacks proper guidance for intra- and cross-modality integration, especially when dealing with multiple modalities. The proposed MPT addresses this challenge by providing modality-specific guidance for aligning multimodal information within MFA-ViT. Unlike SPT, MPT establishes a unified bridge for three modalities, intertwining input sequences while promoting cross-modality interaction while preserving their distinct characteristics.

To sum up, we leverage the MFA and MPT within the ViT architecture to address FBR problems. These mechanisms are devised for effective multiple modalities alignment, especially to accentuate the common distinctive features of the face and periocular, with a specific focus on the ocular region, as they play a pivotal role in achieving exceptional performance in intra- and cross-modality recognition tasks. Additionally, our approach derives advantages from incorporating soft-biometric attributes as supplementary information.

The contributions of this paper are summarized as follows:

- A novel FBR framework is introduced to address intra- and cross-modality recognition by integrating face and periocular modalities with soft-biometric attributes.
- A MFA-ViT is designed to substantiate the FBR notion, enabling the effective fusion of three modalities and systematically examining the inter-dependencies between intra- and cross-modality relationships while establishing a modality-invariant embedding to represent identities.
- The MPT mechanism is proposed, crucially guiding the integration of multimodal data to produce rich, coherent embeddings, capturing intricate relationships between different biometric modalities.

2. Related Work

Conditional biometrics has garnered attention as a promising solution to address the limitations of multimodal biometrics [8, 9, 11]. These studies illustrate the performance improvements achieved by conditioning the face modality with soft-biometrics, particularly in challenging environmental variations and occlusions. A previous study has utilized knowledge distillation techniques [12] to enhance periocular modality performance by incorporating face biometrics. In a similar vein, [22] introduces an approach for face-conditioned periocular recognition. This approach employs a conditional biometrics contrastive loss within a shared-parameter convolutional network.

Nevertheless, FBR offers a more robust solution due to its inherent flexibility to handle intra- and cross-modality recognition tasks without specific conditioning, making it a better choice for demanding real-world applications.

Cross-modality biometrics. Prior studies in cross-modality biometrics have primarily focused on face-voice

recognition [3, 13, 20, 36], while others like [5, 34] focused on bridging visible light face images with alternative modalities such as infrared or depth images. Most recently, [33] introduces HA-ViT, utilizing a face-periocular contrastive learning approach for cross-modality recognition. This approach effectively demonstrates how contrastive learning can proficiently align and differentiate these modalities, enhancing cross-modality recognition task performance. However, previous works tend to neglect the wider utility, especially in tasks related to intra-modality recognition. Furthermore, while the importance of face and periocular biometrics in cross-modality recognition is acknowledged, effectively capturing their intricate relationships remains challenging. [5, 22, 33].

Our approach goes beyond the boundaries of cross-modality biometric recognition by harnessing the MFA and MPT modules within the ViT architecture. Supplementary Material Section 8.1 highlights that these modules emphasize the shared salient features of the face and periocular, particularly the eyes, which are crucial for excelling in intra- and cross-modality recognition tasks.

Prompt Tuning involves generating task-specific continuous vectors using gradient descent [16]. These vectors are designed to guide a pre-trained transformer model to perform specific tasks without requiring extensive fine-tuning of the entire model. This technique has recently been extended to image classification tasks, as explored by [10, 39]. In this context, the learnable visual prompt tuning (VPT) is used to adapt pre-trained ViT models, resulting in improved performance compared to conventional fine-tuning.

However, applying VPT directly to FBR poses challenges. These challenges arise from the inherent limitations of guiding alignment within VPT methodologies, mainly when dealing with the complexities of multiple modalities. To this end, our proposed MPT addresses the alignment of multiple modality embeddings within a shared embedding space that captures multimodal information. Importantly, this unified alignment is a novel contribution not explored in existing literature.

3. Methodology

3.1. Network Architecture

MFA-ViT is built upon a shared-parameter network architecture that accommodates input from three sources: face ($\mathbf{I}_f$), periocular ($\mathbf{I}_p$), and soft-biometric attributes ($\mathbf{I}_a$), as illustrated in Figure 2. The MFA-ViT is based on the ViT architecture [4] due to its effectiveness in handling multi-modality fusion without explicit modifications [38]. Further justifications are provided in Supplementary Material Section 8.3.

Given a pair of image patches $\mathbf{I}_f$ and $\mathbf{I}_p$, each is tokenized to yield embeddings $\mathbf{Z}_f$ and $\mathbf{Z}_p$ with dimension d, and it is set to $1,024$. To incorporate $\mathbf{I}_a$ into the network, we utilize feature tokenizer [7], in order to transform the input $\mathbf{I}_a$ into embeddings $\mathbf{Z}_a \in \mathbb{R}^{1\times d}$. Feature tokenizer enables the seamless integration of categorical-based soft-biometric attributes with image-based face and periocular modalities.

The network takes in biometric token embedding $\mathbf{Z}_*$ where $*$ denotes f, p, or a, a learnable class token embedding $\mathbf{T}_*$ as well as a learnable prompt token embedding $\mathbf{P}_*$. Subsequently, these embeddings are directed to the multimodal fusion attention (MFA) block, B_m where $m = 1, \cdots, M$. Each MFA block comprises MFA layers F_n with $n = 0, 1, \cdots, N-1$. Each F_n is constructed by a 3×3 depth-wise Conv (DWS-Conv) layer, a depth-wise fusion Conv-based multi-head self-attention (DWFC-MSA) layer, a 1×1 Conv layer, and a LeakyReLu (R_{Leaky}) activation layer. In this paper, $M = 2$ and $N = 4$ are adopted.

The embeddings $\mathbf{T}_*$, $\mathbf{Z}_*$, and $\mathbf{P}_*$ are concatenated and subsequently fed into the F_n layer within each B_m, which is outlined as follows:

$$F_{n+1}(\mathbf{K}_{*,n}) = R_{\text{Leaky}}(\text{Conv}([\text{DWS-Conv}(\mathbf{K}_{*,n}), \text{DWFC-MSA}(\mathbf{K}_{*,n})])) + \mathbf{K}_{*,n}, \quad (1)$$

where $\mathbf{K}_{*,n} = [\mathbf{T}_{*,n}, \mathbf{Z}_{*,n}, \mathbf{P}_{*,n}] \in \mathbb{R}^{S\times H\times W}$ and S, H, W denote the number of input embeddings, height, width, respectively. The DWS-Conv layer specializes in extracting distinct local features from the $\mathbf{I}_f$ and $\mathbf{I}_p$ patches while simultaneously considering the associations encoded by the $\mathbf{I}_a$. These associations substantiate the learning of multi-modal embeddings, particularly facilitated by the $\mathbf{P}_*$.

The DWFC-MSA layer is tailored to capture the relationships within and across modalities within the $\mathbf{I}_f$ and $\mathbf{I}_p$ patches. The presence of $\mathbf{I}_a$ enriches this layer, allowing it to achieve a holistic comprehension. The DWFC-MSA layer collaborates seamlessly with $\mathbf{P}_*$ to support the development of this understanding. The DWFC-MSA (E) layer is structured as follows:

$$E'_{n+1}(\mathbf{K}_{*,n}) = \text{C-MSA}(\text{Norm}(\mathbf{K}_{*,n})) + \mathbf{K}_{*_n}, \quad (2)$$

$$E_{n+1}(\mathbf{K}_{*,n}) = \text{MLP}(\text{Norm}(E'_{n+1}(\mathbf{K}_{*,n}))) + E'_{n+1}(\mathbf{K}_{*,n}), \quad (3)$$

where $\text{Norm}(\cdot)$ denotes layer normalization, $\text{C-MSA}(\cdot)$ refers to 3×3 depth-wise Conv-based MSA layer, and $\text{MLP}(\cdot)$ represents multi-layer perception. Noted that the input tokens for DWS-Conv and DWFC-MSA layers are reshaped for spatial dimensions.

As the final step, the network encodes joint embeddings $\mathbf{J}_*$ by aggregating the MPT embeddings $\mathbf{P}'_{*,N}$ from each

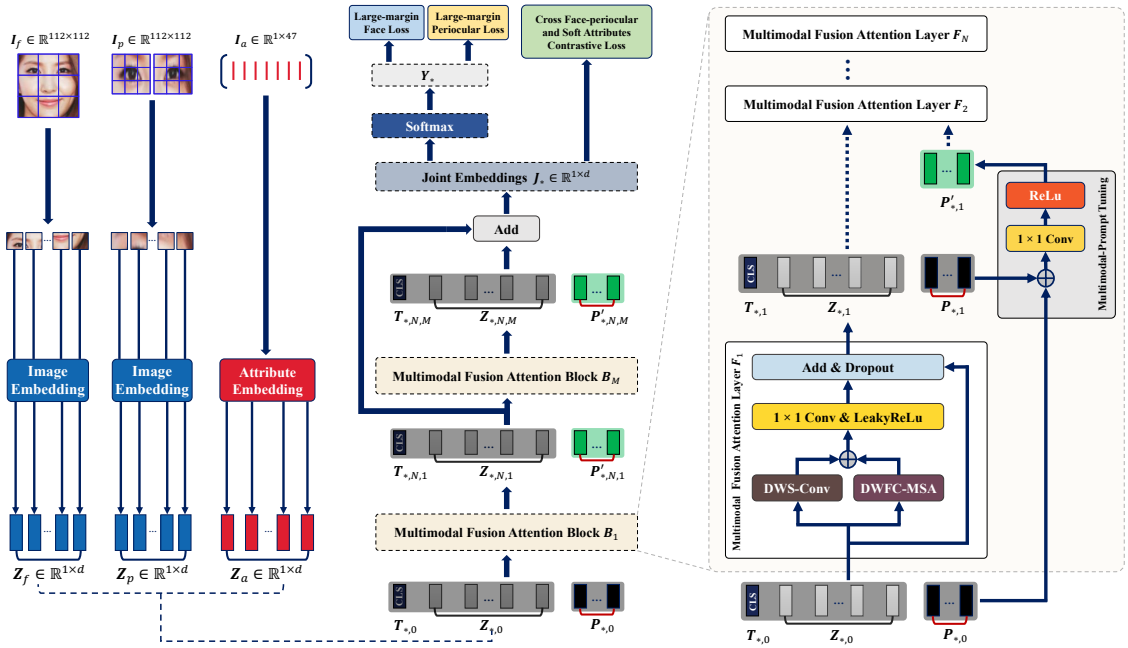

Figure 2. Network architecture of Multimodal Fusion Attention (MFA) Vision Transformer with Multimodal-Prompt Tuning (MPT).

B_m via addition and followed by an average pooling (Avgpool) operation. For classification, we utilize $\mathbf{J}_*$ as the input to the softmax layer, resulting in the softmax vector $\mathbf{Y}_*$. We opt to employ $\mathbf{J}_*$ instead of a class token $\mathbf{T}_*$ as commonly utilized in ViT, is driven by the hypothesis that the MPT, explicitly designed for the multimodal fusion task, can offer a more effective way to capture and leverage cross-modality information. We explore the impacts of this option in Section 4.4.4. The final formulation is defined as follows:

$$\mathbf{J}_* = \text{Avgpool}(\sum_{m=1}^{M} \mathbf{P}'_{*,N,M}), \tag{4}$$

$$\mathbf{Y}_* = \text{Softmax}(\mathbf{J}_*). \tag{5}$$

3.2. Multimodal-Prompt Tuning

MPT is incorporated at the input space after the embedding layers, which are attached to $\mathbf{T}_*$ and $\mathbf{Z}_*$, as illustrated in Figure 2. The MPT embeddings ($\mathbf{P}'_*$) play a pivotal role in guiding the process of multimodal feature prompt learning in each F_n layer. Utilizing this guidance enables a subtle and detailed exploration of relationships among the diverse modalities and attributes under consideration.

Specifically, MPT is integrated at the input space of each layer in B_m. The structure of MPT involves a 1×1 Conv layer, followed by a ReLU layer, applied to $\mathbf{P}_{*,n}$ embeddings. We designate the set of learnable $\mathbf{P}'_{*,n}$ embeddings with dimension size of d. Section 4.4.2 emphasizes MPT embeddings' role in discerning intricate details within multimodal features, with Figure 5 demonstrating their efficacy in capturing eye regions across facial and periocular images. Additional studies can be found in Supplementary Material Section 8.2. The input space with MPT embeddings can be computed as:

$$[\mathbf{T}_{*,n+1}, \mathbf{Z}_{*,n+1}, \mathbf{P}'_{*,n+1}] = F_{n+1}([\mathbf{T}_{*,n+1}, \mathbf{Z}_{*,n+1}, L_{n+1}(\mathbf{P}_{*,n}, \mathbf{P}_{*,n+1})]), \tag{6}$$

where $\mathbf{P}_{*,n}$ and $\mathbf{P}_{*,n+1}$ represent the previous and current input prompt embeddings. L_{n+1} is calculated using a ReLU activation applied to the output of 1×1 Conv as defined by:

$$L_{n+1}(\mathbf{P}_{*,n}, \mathbf{P}_{*,n+1}) = \text{ReLu}(\text{Conv}([\mathbf{P}_{*,n}, \mathbf{P}_{*,n+1}])). \tag{7}$$

3.3. Multimodal Contrastive Loss Function

For training, we employ a dual strategy that combines both large-margin softmax loss ($\mathcal{L}_{\text{LM}}$) [14] and contrastive loss ($\mathcal{L}_{\text{CL}}$) [33]. This strategy aims to learn intra-modality relationships within face (f) and periocular (p), and cross-modality relationships encompassing f, p, and soft-biometric attributes (a).

Specifically, the $\mathcal{L}_{\text{LM}}$ contributes to shaping the embedding space to better discriminate between different identities, which is essential for intra-modality recognition. The $\mathcal{L}_{\text{LM}}$ is given as follows:

$$\mathcal{L}_{\text{LM}_\ddagger} = -\log(\Phi_\ddagger) + \frac{\lambda}{2} \sum_{c \neq l} \left[\Psi_{\ddagger,c} - \frac{1}{C-1} \cdot \log(\Psi_{\ddagger,c}) \right], \tag{8}$$

where Φ denotes softmax function, $\Psi_{\ddagger,c}$ is to maximize the margin of discriminated embedding between the identity of the *same modalities*, C is the number of identities, λ is to control the degree of degrading labels, and l is the label. λ is set to 0.3 same as [14]. The $\mathcal{L}_{\text{LM},\ddagger}$ is computed individually to f and p modalities, with $\ddagger$ representing either f or p.

On the other hand, the $\mathcal{L}_{\text{CL}}$ accentuates the embedding of the same identity samples, fostering closer proximity in the embedding space while simultaneously pushing apart samples from different identities across three modalities, i.e., f, p, and a. The loss function is implemented following the formulation introduced by [33] and [40] to establish significant connections between cross-modality relationships. The $\mathcal{L}_{\text{CL}}$ can be expressed as:

$$\mathcal{L}_{\text{CL}} = \mathcal{L}_{\text{CM}}(f,p) + \mathcal{L}_{\text{CM}}(f,a) + \mathcal{L}_{\text{CM}}(p,a), \tag{9}$$

$$\mathcal{L}_{\text{CM}}(x_u, y_u) = - \log \frac{\sigma(x_u, y_u)}{\sigma(x_u, y_u) + \alpha\left(\delta(x_u, x_v)\right) + \delta(x_u, y_v)}, \tag{10}$$

where $\sigma(x_u, y_u) = \exp(\frac{\mathbf{J}_{x_u}^T \mathbf{J}_{y_u}}{\theta})$ serves to map the embeddings $\mathbf{J}$, of *distinct modalities* x and y, yet sharing the *same identity* u, into a shared embedding space. The hyperparameter θ is introduced to extend the range of $\mathbf{J}_{x_u}^T \mathbf{J}_{y_u}$ to facilitate the model's convergence.

Furthermore, $\delta(x_u, x_v) = \sum_{x_v \in \mathbf{N}_u^X} \sigma(x_u, x_v)$ represents pairs of data samples sharing the same modality but differing in identity v. Here, $x_v \in \mathbf{N}_u^X$ designates *intra-modality pairs*, which are pairs of *different samples* within the *same modality*. Similarly, $\delta(x_u, y_v) = \sum_{y_u \in \mathbf{N}_v^Y} \sigma(x_u, y_v)$ characterizes pairs of data samples from *different modalities* and *identities*, thereby ensuring they remain distinguishable in the shared embedding space. Here, $y_v \in \mathbf{N}_u^Y$ refers to *cross-modality pairs* between *different samples*. Leveraging $\delta(x_u, x_v)$ and $\delta(x_u, y_v)$ empowers the model to effectively discern between the high similarities in cross-modalities from different identities.

In our study, α is set to 0.8 same as [40]. Additionally, we set θ to 0.03 for the $\mathcal{L}_{\text{CM}}(f,p)$ term, and to 0.04 for both $\mathcal{L}_{\text{CM}}(f,a)$ and $\mathcal{L}_{\text{CM}}(p,a)$ terms.

The total loss ($\mathcal{L}_{\text{total}}$) is given as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}_f} + \mathcal{L}_{\text{LM}_p} + \mathcal{L}_{\text{CL}}. \tag{11}$$

3.4. Flexible Recognition

To determine a person's identity, the trained model is designed flexibly to accept inputs from the f or p modalities. Specifically, the softmax layer is detached for recognition, and the modality-invariant embedding is extracted from $\mathbf{J}_\ddagger$, where $\ddagger$ denotes f or p. The identity of $\mathbf{J}_\ddagger$ can be decided based on

$$\psi = \max_k \left[s(\mathbf{G}_{k,\ddagger}, \mathbf{J}_\ddagger)\right] \tag{12}$$

where $s(\cdot)$ calculates a similarity score between the unknown identity $\mathbf{J}_\ddagger$ and the gallery sets $\mathbf{G}_{k,\ddagger}$, k refers to the number of identities in gallery set.

4. Experiment

4.1. Dataset

4.1.1 Training dataset

The training dataset comprises modalities $\mathbf{I}_f$, $\mathbf{I}_p$, and $\mathbf{I}_a$, which are sampled from the VGGFace2 dataset [2] and the MAAD-Face dataset [31]. After a comprehensive dataset review, we have selected 1.49 million samples with 9,131 identities. For each identity, we have paired face $\mathbf{I}_f$ and periocular $\mathbf{I}_p$ images with 47 attributes $\mathbf{I}_a$ representing the identities, detailed in [31]. We randomly partitioned the identities, allocating them into training and validation sets in an 80:20 ratio.

4.1.2 Evaluation dataset

To have a fair comparison, we have selected four public datasets, namely Ethnic [32], FaceScrub [21], IMDB [26], and Cross-Modal DB [33]. Further details can be found in Supplementary Material Section 7. These datasets are benchmarks for evaluating our network's intra- and cross-modality matching performance. We adhere to the protocol in [32], which involves matching a probe from the gallery sets. In this evaluation, all trained models serve as feature extractors for $\mathbf{I}_f$ and $\mathbf{I}_p$ across both gallery and probe sets. The matching process is executed using cosine similarity.

4.2. Experimental Setup

MFA-ViT is trained over 50 epochs with a batch size of 64. The input sizes for $\mathbf{I}_f$ and $\mathbf{I}_p$ are both set to $3 \times 112 \times 112$, while $\mathbf{I}_a$ is 1×47. Each $\mathbf{I}_f$ and $\mathbf{I}_p$ image is divided into 14 patches, and the size of each patch is 8×8. We minimize the total loss using the AdamW Optimizer [17]. We employ a learning rate of 1e-4, a weight decay parameter 1e-5, and a dropout rate of 0.1.

4.3. Experimental Results

To assess the FBR performance, we conduct a comprehensive comparison by re-implementing a baseline model trained solely on $\mathbf{I}_f$ and another solely on $\mathbf{I}_p$. Furthermore, we examined several relevant models designed for intra- and cross-modality recognition problems, i.e., [32], [22], and [5]. We also aim to provide a broader comparison encompassing recent studies such as HA-ViT [33]

and ViT/VPT [10] to examine the effectiveness of our proposed model. Noted that we have re-implemented these models to adapt input embeddings, customizing them for fair comparisons, while adhering to the same experimental settings and protocols. In Table 1, we present the performance comparisons for intra- and cross-modality recognition tasks, with the primary metric being rank-1 recognition accuracy. Specifically, the cross-modality recognition tasks encompass scenarios such as face gallery vs. periocular probe (*f–p*) and periocular gallery vs. face probe (*p–f*), while intra-modality recognition focuses on face gallery vs. face probe (*f–f*) and periocular gallery vs. periocular probe (*p–p*).

4.3.1 Intra-modality Recognition

As indicated in Table 1, the MFA-ViT/MPT model exhibits exceptional recognition performance on both the Ethnic and FaceScrub datasets, achieving high accuracy rates of 94.82% and 95.71% for the *f–f* and 89.98% and 93.06% for the *p–p*, respectively. Furthermore, even when confronted with more challenging datasets such as IMDB and Cross-Modal DB, MFA-ViT/MPT maintains its competitive advantage, yielding mean accuracy rates of 86.03% and 85.88% for *f–f* and 80.53% and 76.54% for *p–p*. These outcomes underscore the model's consistent outperformance of baseline and competing models.

The performance of both the baseline and [32] is suboptimal, primarily attributed to their extensive dependence on the individual $\mathbf{I}_f$ and $\mathbf{I}_p$ modalities. In contrast, the models presented by [5] and [22] have exhibited satisfactory results by adeptly capturing adaptive relational knowledge between $\mathbf{I}_f$ and $\mathbf{I}_p$ embeddings. Nevertheless, the necessity for both face and periocular inputs in these models restricts their adaptability.

Notably, the experiments reveal that models with prompt tuning, such as ViT/VPT, generally outperform those without, highlighting the efficacy of the prompt tuning. However, ViT/VPT still exhibits a marginal performance gap compared to our model. This observation justifies that the MFA-ViT and MPT are pivotal in exploring complex relationships among the modalities.

4.3.2 Cross-modality Recognition

In Table 1, the MFA-ViT/MPT model exhibits notable average recognition accuracies for both the *f–p* and *p–f* scenarios. Specifically, in the *f–p*, the model achieved accuracies of 86.70% on the Ethnic dataset, 90.38% on FaceScrub, 75.28% on IMDB, and 72.01% on Cross-Modal DB. Conversely, in the *p–f* configuration, it attains impressive average accuracies of 89.06% on Ethnic, 92.02% on FaceScrub, 76.36% on IMDB, and 75.96% on Cross-Modal DB. The outcomes underscore the consistent superiority of our model over existing methods, highlighting its effectiveness in addressing cross-modality recognition across four datasets.

The baseline among benchmark methods significantly underperformed, scoring below 1% in evaluations, indicating the distinct nature of face and periocular modalities despite periocular being a face subset. Hence, modality alignment is crucial for cross-modal matching. In contrast, [5] and [22] employed contrastive learning for enhanced performance, introducing a trade-off between intra- and cross-modality recognition. As intra-modality performance improved, cross-modality recognition degraded, and vice versa, compared to our model. Their focus on broad embedding space alignment overshadowed nuances in $\mathbf{I}_f$ and $\mathbf{I}_p$ attributes, potentially causing suboptimal recognition across modalities.

As shown in Table 1, ViT/VPT and HA-ViT, even trained with $\mathbf{I}_a$, did not perform as effectively as our approach. This can be attributed to HA-ViT primarily focusing on utilizing cross-modality loss functions, neglecting the benefits of prompt tuning for enhancing multimodal features. On the other hand, ViT/VPT, while utilizing prompt tuning, encountered challenges in effectively handling multimodal features.

In summary, results demonstrate MFA adeptly aligns cross-modality relations, enabling precise matching without contrastive learning trade-offs. Simultaneously, MPT efficiently guides the model, capturing cross-modality and multimodal dependencies. Collaborative synergy between MPT and $\mathbf{I}_a$ enhances the understanding, contributing to superior cross-modal recognition. This underscores the efficacy of our approach to FBR challenges.

4.4. Ablation Study

4.4.1 Effects of Soft-biometric Attributes

We undertook an ablation study to evaluate the effectiveness of $\mathbf{I}_a$ using MFA-ViT and HA-ViT. Notably, both networks were trained to employ the MPT approach. For a fair comparison, soft-biometric attributes are integrated into the input embeddings of HA-ViT with a feature tokenizer during training. Table 1 and Figure 3 demonstrate that MFA-ViT outperforms other models when equipped with $\mathbf{I}_a$. Surprisingly, HA-ViT, whether with or without $\mathbf{I}_a$, lagged behind MFA-ViT, even though both models employed the identical $\mathcal{L}_{\text{total}}$ loss function.

One plausible rationale behind the enhanced utilization of $\mathbf{I}_a$ by MFA-ViT lies in its architectural refinement, designed for seamless integration of features derived from diverse modalities with $\mathbf{I}_a$. This empowers the network to process multiple information sources efficiently, equipping it to capture the rich information encapsulated within $\mathbf{I}_a$. In contrast, HA-ViT, despite sharing a similar network structure, appears to lack an effective fusion mechanism, rely-

Table 1. FBR performance comparisons on intra-modality (*f–f* and *p–p*) and cross-modality (*f–p* and *p–f*) recognition tasks in terms of rank-1 recognition (%). The best accuracy is written in bold.

Model	Ethnic				FaceScrub				IMDB				Cross-Modal DB			
	f–f	*p–p*	*f–p*	*p–f*	*f–f*	*p–p*	*f–p*	*p–f*	*f–f*	*p–p*	*f–p*	*p–f*	*f–f*	*p–p*	*f–p*	*p–f*
Model trained independently on Face or Periocular																
Baseline	80.61	61.79	0.39	0.33	76.10	63.29	0.47	0.42	62.08	55.31	0.10	0.09	60.43	51.43	0.09	0.07
Model trained with Face and Periocular																
Tiong *et al.* [32]	80.30	65.42	0.44	0.51	76.46	72.13	0.57	0.61	65.31	54.91	0.18	0.19	64.70	50.18	0.11	0.13
George *et al.* [5]	85.98	66.90	54.18	59.46	89.80	77.76	62.31	66.82	72.58	54.72	34.91	43.57	70.56	47.16	29.50	38.72
Ng *et al.* [22]	92.55	79.29	71.54	75.36	91.82	85.74	75.15	78.36	77.19	71.99	58.56	61.19	76.21	65.63	58.13	61.35
ViT/VPT [10]	91.80	83.94	76.68	77.74	93.11	89.26	83.35	85.17	78.57	72.15	60.46	62.32	80.97	69.01	63.13	66.42
HA-ViT [33]	91.36	84.32	76.34	77.65	92.13	88.52	80.27	81.80	78.23	71.92	59.49	59.68	78.76	66.87	59.93	61.87
Model trained with Face, Periocular, and Soft-biometric Attributes																
ViT/VPT [10]	92.95	85.68	83.37	86.10	93.57	90.84	87.43	88.61	81.97	74.59	69.40	70.21	82.76	70.92	65.99	69.11
HA-ViT [33]	91.72	85.10	80.03	81.61	92.46	88.70	84.33	85.63	78.81	71.42	64.13	65.49	78.81	67.22	62.34	64.03
MFA-ViT/MPT	**94.82**	**89.98**	**86.70**	**89.07**	**95.71**	**93.06**	**90.38**	**92.02**	**86.03**	**80.53**	**75.28**	**76.36**	**85.88**	**76.54**	**72.01**	**75.96**

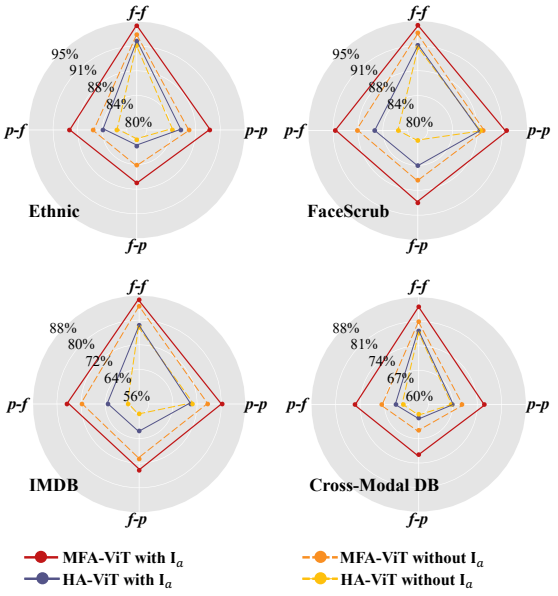

Figure 3. Performance comparison on the proposed model trained with or without $\mathbf{I}_a$ against HA-ViT.

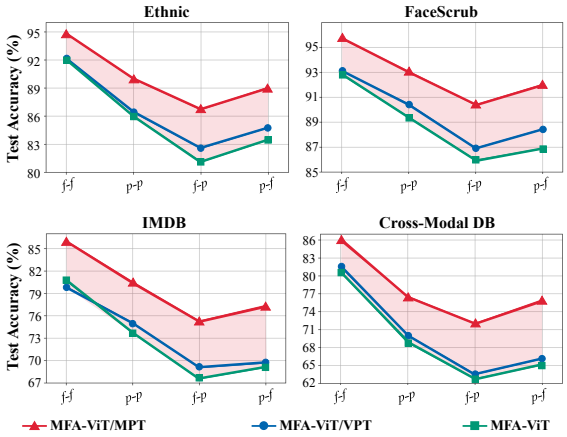

Figure 4. Performance comparison on MPT, VPT, and no prompt. This analysis exclusively focuses on the evaluation of MFA-ViT trained using $\mathbf{I}_f$, $\mathbf{I}_p$, and $\mathbf{I}_a$.

ing primarily on simple concatenation for aggregating multimodal features. This primitive fusion approach in HA-ViT may account for its comparatively suboptimal utilization of $\mathbf{I}_a$ compared to MFA-ViT.

4.4.2 With and Without MPT

To gauge the efficacy of the MPT strategy, we compare the MFA-ViT's performance with and without the incorporation of MPT. As illustrated in Figure 4, a progressive enhancement in performance is observed when MFA-ViT is integrated with MPT, compared to its operation without MPT. However, it is crucial to underscore that the performance of MFA-ViT without MPT lags in delivering substantial rank-1 results in intra- and cross-modality recognition tasks.

To gain deeper insights into the advantages of the MPT strategy, we present visualization results using the Grad-CAM [28]. As shown in Figure 5, MFA-ViT/MPT focuses intensively on the shared areas of the eye regions in the $\mathbf{I}_f$ and $\mathbf{I}_p$ images. Conversely, MFA-ViT does not employ a prompt strategy, further limiting the model's ability to discern less specific features compared to the comprehensive approach offered by the MPT strategy.

These observations vividly showcase the robustness of MPT. Without the guidance of MPT, MFA-ViT struggles to discern and accentuate specific features across modalities precisely. However, the integration of MPT enables MFA-ViT to manage input sequences adeptly across its multimodal fusion attention layers. This goes beyond just aligning different modalities; it is also about discerning and preserving the unique attributes of each while fostering cohesive cross-modality collaboration.

Table 2. Performance comparisons on classification head inputs (CLS and PRM) in terms of rank-1 recognition (%). The best accuracy is indicated in bold.

Head Input		Ethnic				FaceScrub				IMDB				Cross-Modal DB			
CLS	*PRM*	*f–f*	*p–p*	*f–p*	*p–f*	*f–f*	*p–p*	*f–p*	*p–f*	*f–f*	*p–p*	*f–p*	*p–f*	*f–f*	*p–p*	*f–p*	*p–f*
✓		94.57	89.18	86.24	88.34	95.26	92.63	89.81	91.51	85.32	79.43	74.52	76.36	85.11	75.53	71.06	74.35
	✓	**94.82**	**89.98**	**86.70**	**89.07**	**95.71**	**93.06**	**90.38**	**92.02**	**86.03**	**80.53**	**75.28**	**77.37**	**85.88**	**76.54**	**72.01**	**75.96**

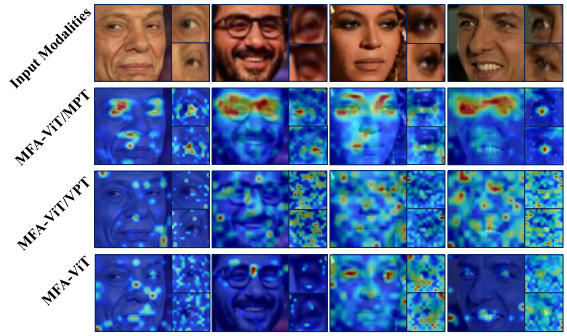

Figure 5. Visualization of activation maps for MFA-ViT with MPT, VPT, and no prompt strategies.

4.4.3 MPT vs VPT

We further investigate the effectiveness of MPT compared to VPT within the MFA-ViT architecture. Since VPT was initially designed only for ViT, we extend it to demonstrate the performance of MPT in utilizing attributes in contrast to VPT. As illustrated in Figure 4, we observe a gradual increase in the performance of MFA-ViT/MPT compared to MFA-ViT/VPT. However, it is essential to note that the performance of MFA-ViT/VPT still falls short of achieving significant rank-1 results in intra- and cross-modality recognition tasks.

Delving deeper, the visualizations in Figure 5 elucidate that MFA-ViT/MPT is impressive because it adeptly attends the ocular region on the $\mathbf{I}_f$ image and simultaneously on the $\mathbf{I}_p$ modalities. In contrast, MFA-ViT/VPT appears to have difficulty differentiating specific features between $\mathbf{I}_f$ and $\mathbf{I}_p$ modalities. This is because the VPT strategy was initially designed for task-specific features, lacking guidance for the model to understand the intricate association between multimodal features.

These findings underscore the prowess of MPT in facilitating a seamless interplay between $\mathbf{I}_f$ and $\mathbf{I}_p$ modalities with $\mathbf{I}_a$. The nuanced guidance steers the model towards a delicate understanding of the modalities, distinguishing it from VPT. Moreover, the results indicate that VPT's functionality is impacted by neglecting important multimodal features within each layer of the learning process.

4.4.4 Impacts of Classification Head Inputs

In this subsection, we investigate the influence of different classification head inputs on our model's performance in FBR. We consider two types of classification head input: the class token embedding $\mathbf{T}_*$ (*CLS*) follow the standard practice as in ViT [4] and the joint embeddings $\mathbf{J}_*$ from multimodal-prompt token embedding (*PRM*). As shown in Table 2, the result reveals a significant performance degradation when the model is trained with *CLS*. In contrast, training with *PRM* consistently improves performance across all benchmark datasets.

The superiority of *PRM* over *CLS* is underscored by these findings. *PRM* exhibits effective alignment with FBR tasks, enabling the model to acquire a more discriminative embedding. This is achieved through prompt embeddings to guide attention toward specific features and relationships within multimodal data. In contrast, *CLS* lacks cross-modality feature alignment, potentially leading to a failure in capturing delicate relationships crucial for addressing FBR problems.

5. Discussion and Conclusion

This paper introduces the MFA-ViT approach to address the intricate challenges of flexible biometric recognition (FBR). MFA-ViT leverages an MFA and an MPT to capture cross-modality and multimodal dependencies in embedding. The MPT is crucial in facilitating the acquisition of intermodal associations, which is vital in intra- and cross-modality recognition tasks. The integration enhances the comprehensive understanding of data, particularly in challenging real-world scenarios, significantly boosting recognition performance. Additionally, incorporating soft-biometric attributes provides further contextual insights, strengthening the discriminative potential of our embeddings. For future work, we see the potential to extend FBR to encompass other biometric modalities, paving the way for exceptional accuracy and efficiency in diverse recognition tasks.

6. Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NO. NRF-2022R1A2C1010710).

References

[1] Elisa Barroso, Gil Santos, Luis Cardoso, Chandrashekhar Padole, and Hugo Proença. Periocular recognition: How much facial expressions affect performance? *Pattern Anal. Appl.*, 19:517–530, 2016.

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Int. Conf. Autom. Face Gesture Recog.*, pages 67–74, 2018.

[3] Kai Cheng, Xin Liu, Yiu-ming Cheung, Rui Wang, Xing Xu, and Bineng Zhong. Hearing like seeing: Improving voice-face interactions and associations via adversarial deep semantic matching network. In *ACM MM*, page 448–455, 2020.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[5] Anjith George and Sebastien Marcel. Cross modal focal loss for RGBD face anti-spoofing. In *CVPR*, pages 7878–7887, 2021.

[6] Ester Gonzalez-Sosa, Julian Fierrez, Ruben Vera-Rodriguez, and Fernando Alonso-Fernandez. Facial soft biometrics for recognition in the wild: Recent works, annotation, and cots evaluation. *IEEE Trans. Inf. Forensics Secur.*, 13(8):2001–2014, 2018.

[7] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *NeurIPS*, pages 18932–18943, 2021.

[8] JongWon Hwang, Leslie Ching Ow Tiong, and Andrew Beng Jin Teoh. Towards face representation learning conditioned on the soft biometrics. In *Int. Conf. Mach. Vis. Appl.*, page 1–7, 2022.

[9] Seyed Mehdi Iranmanesh, Ali Dabouei, and Nasser Nasrabadi. Attribute adaptive margin softmax loss using privileged information. In *BMVC*, pages 1–13, 2020.

[10] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, page 709–727, 2022.

[11] Luo Jiang, Juyong Zhang, and Bailin Deng. Robust RGB-D face recognition using attribute-aware loss. *IEEE TPAMI*, 42 (10):2552–2566, 2020.

[12] Yoon Gyo Jung, Cheng Yaw Low, Jaewoo Park, and Andrew Beng Jin Teoh. Periocular recognition in the wild with generalized label smoothing regularization. *IEEE Sign. Process. Letters*, 27:1455–1459, 2020.

[13] Changil Kim, Hijung Valentina Shin, Tae-Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik. On learning associations of faces and voices. In *ACCV*, pages 276–292, 2018.

[14] Takumi Kobayashi. Large margin in softmax cross-entropy loss. In *BMVC*, pages 1–12, 2019.

[15] Nagashri N Lakshminarayana, Nishant Sankaran, Srirangaraj Setlur, and Venu Govindaraju. Multimodal deep feature aggregation for facial action unit recognition using visible images and physiological signals. In *Int. Conf. Autom. Face Gesture Recog.*, pages 1–4, 2019.

[16] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):1–35, 2023.

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, pages 1–10, 2019.

[18] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *Conf. Graph. Patterns Images*, pages 471–478, 2018.

[19] Shervin Minaee, Amirali Abdolrashidi, Hang Su, Mohammed Bennamoun, and David Zhang. Biometrics recognition using deep learning: A survey. *Artif. Intell. Rev.*, 56 (8):8647–8695, 2023.

[20] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable PINS: Cross-modal embeddings for person identity. In *ECCV*, page 71–88, 2018.

[21] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *ICIP*, pages 343–347, 2014.

[22] Tiong-Sik Ng, Cheng-Yaw Low, Jacky Chen Long Chai, and Andrew Beng Jin Teoh. Conditional multimodal biometrics embedding learning for periocular and face in the wild. In *ICPR*, pages 812–818, 2022.

[23] Chandrashekhar N. Padole and Hugo Proença. Periocular recognition: Analysis of performance degradation factors. In *Int. Conf. Biometrics (ICB)*, pages 439–445, 2012.

[24] Unsang Park, Arun Ross, and Anil K. Jain. Periocular biometrics in the visible spectrum: A feasibility study. In *Int. Conf. Biometrics Theory Appl. Syst. (BTAS)*, pages 1–6, 2009.

[25] Hugo Proença and João C. Neves. Deep-PRWIS: Periocular recognition without the iris and sclera using deep learning frameworks. *IEEE Trans. Inf. Forensics Secur.*, 13(4):888–896, 2018.

[26] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 126(2):144–157, 2018.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018.

[28] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, pages 1–14, 2015.

[30] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Int. Conf. Mach. Learn. (ICML)*, page 6105–6114, 2019.

[31] Philipp Terhörst, Daniel Fährmann, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Maadface: A massively annotated attribute dataset for face images. *IEEE Trans. Inf. Forensics Secur.*, 16:3942–3957, 2021.

[32] Leslie Ching Ow Tiong, Seong Tae Kim, and Yong Man Ro. Multimodal facial biometrics recognition: Dual-stream convolutional neural networks with multi-feature fusion layers. *Image Vis. Comput.*, 102:103977, 2020.

[33] Leslie Ching Ow Tiong, Dick Sigmund, and Andrew Beng Jin Teoh. Face-periocular cross-identification via contrastive hybrid attention vision transformer. *IEEE Sign. Process. Letters*, 30:254–258, 2023.

[34] Hanrui Wang, Xingbo Dong, Zhe Jin, Jean-Luc Dugelay, and Massimo Tistarelli. Cross-spectrum face recognition using subspace projection hashing. In *ICPR*, pages 615–622, 2021.

[35] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.

[36] Rui Wang, Xin Liu, Yiu-ming Cheung, Kai Cheng, Nannan Wang, and Wentao Fan. Learning discriminative joint embeddings for efficient face and voice association. In *ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, page 1881–1884, 2020.

[37] Gou Wei, Li Jian, and Sun Mo. Multimodal(audio, facial and gesture) based emotion recognition challenge. In *Int. Conf. Autom. Face Gesture Recognit.*, pages 908–911, 2020.

[38] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. *IEEE TPAMI*, 45(10): 12113–12132, 2023.

[39] Yuhao Zhu, Min Ren, Hui Jing, Linlin Dai, Zhenan Sun, and Ping Li. Joint holistic and masked face recognition. *IEEE Trans. Inf. Forensics Secur.*, 18:3388–3400, 2023.

[40] M. Zolfaghari, Y. Zhu, P. Gehler, and T. Brox. CrossCLR: Cross-modal contrastive learning for multi-modal video representations. In *ICCV*, pages 1430–1439, 2021.