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

Supplementary Material

Table 3. Summary of Evaluation Datasets.

	Ethnic	FaceScrub	IMDB	Cross-Modal DB
No. Identity	329	530	2,129	2,239
No. Modality Pair	25,816	79,876	89,424	190,876
Ratio of Gallery:Probe	5:95	40:60	40:60	40:60
No. Gallery:Probe Sets	1	3	3	3
Diversity of Ethnicity	Yes	N/A	Yes	Yes

N/A refers to not available.

7. Details of Evaluation Protocols

Four publicly accessible datasets are selected in this study, namely, Ethnic [32], FaceScrub [21], IMDB [26], and Cross-Modal DB [33]. Details of these datasets are presented in Table 3. We follow the evaluation protocol outlined in [32], which involves matching of a probe image with images from the gallery sets. For the Ethnic dataset, the ratio of gallery to probe sets is 5:95, while for FaceScrub, IMDB, and Cross-Modal DB, this ratio is standardized at 40:60.

During the evaluation, all models trained in this study function as feature extractors for both $\mathbf{I}_f$ and $\mathbf{I}_p$ modalities across the gallery and probe sets. The process of matching is carried out using cosine similarity as follows:

$$\cos(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{e} \mathbf{A}_i \mathbf{B}_i}{\sqrt{\sum_{i=1}^{e} (\mathbf{A}_i)^2} \sqrt{\sum_{i=1}^{e} (\mathbf{B}_i)^2}}$$

where $\mathbf{A}$ and $\mathbf{B}$ represent the feature embedding vectors cross the gallery and probe sets, respectively, with the dimension of e.

8. Additional Results

8.1. Analysis on Shared Salient Features

In cross-modality recognition for the face and periocular region, models are expected to focus on the eye region for effective matching. However, as illustrated in Figure 6, it is shown that HA-ViT struggles to align and integrate cross-modal information effectively, missing critical periocular features. In contrast, our approach employs MPT and MFA modules to capture periocular details from facial and periocular images adeptly. This observation underscores the enhanced capability of our method in leveraging salient features for cross-modality biometric recognition.

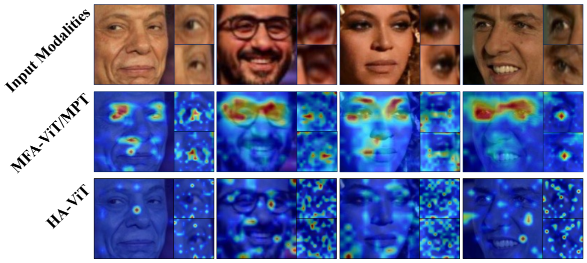

Figure 6. Activation maps of MPT-ViT and HA-ViT

8.2. Impacts of MPT Strategies

We evaluate the effects of two MPT strategies: *Deep* referred explicitly to as MPT-D and *Intermediate* as MPT-I, within our backbone structure MFA-ViT. The MPT-D strategy entails integrating multimodal-prompt token embeddings *throughout the input sequence*, reaching each multimodal fusion attention layer. In contrast, the MPT-I strategy introduces multimodal-prompt token embeddings at *each attention block's output*, resulting in fewer interactions than the MPT-D strategy. This evaluation encompasses intra- and cross-modality recognition and is conducted across four datasets, as summarized in Table 4. All the models are trained with the same number of epochs and hyperparameters.

As highlighted in Table 4, our findings reveal that utilizing the MPT-D strategy consistently improves recognition accuracies for intra- and cross-modality tasks. This improvement can be attributed to the deep integration of prompt embeddings, which allows the model to capture intricate relationships between modalities. However, the performance of MPT-I is closely linked to the depth at which prompt embeddings are inserted. In particular, inserting prompt embeddings at each input of the B_k appears to have a less significant impact on accuracy than MPT-D inserted at every layer.

Furthermore, we also investigated the impact of varying the size of $\mathbf{P}'_*$ and presented the results for sizes 24 and 32 in Table 4. Interestingly, the experiments reveal that a larger $\mathbf{P}'_*$ size (i.e., 32) leads to significantly enhanced model performance. This observation suggests that $\mathbf{P}'_*$ size of 32 consistently delivers reliable performance. Notably, when using a smaller $\mathbf{P}'_*$ size, we observed a decrease in accuracy by at least 1-2% across all tasks.

In addition, the MPT-D strategy significantly contributes to improved accuracies in various recognition tasks, under-

Table 4. Performance comparisons on deep and intermediate MPT and classification head inputs (CLS and PRM) in terms of rank-1 recognition (%). The best accuracy is in bold, and the second-best is in italics.

MPT-D	***MPT-I***	**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*
Using a size of 32 for $\boldsymbol{P'_*}$																			
✓		✓		*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**
	✓	✓		94.40	88.76	85.28	87.55	93.23	90.12	86.29	87.76	83.03	77.49	71.88	72.92	81.35	74.08	70.11	72.29
	✓		✓	94.43	88.76	86.09	87.73	94.82	92.24	88.68	90.12	84.92	78.99	73.05	74.51	84.79	76.03	70.77	73.93
Using a size of 24 for $\boldsymbol{P'_*}$																			
✓		✓		92.73	86.62	83.71	86.01	93.71	89.73	86.29	88.73	81.55	73.90	67.78	70.67	81.54	69.52	64.17	68.68
✓			✓	93.76	87.37	83.35	85.06	94.20	91.13	87.73	89.16	82.92	76.74	70.15	72.05	82.80	72.53	67.14	70.87
	✓	✓		92.54	85.91	82.49	83.79	93.22	89.40	85.54	87.53	80.57	73.42	66.75	69.01	80.80	68.93	63.13	67.37
	✓		✓	92.57	85.58	82.67	83.99	93.49	90.09	87.07	88.67	81.30	74.35	69.07	71.16	81.18	69.88	65.34	69.55

Table 5. FBR performance comparisons on different network backbones 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*
VGGNet-16 [29]	90.16	80.34	61.36	56.44	91.70	88.17	78.84	69.73	76.33	69.95	54.43	46.73	75.67	66.32	50.42	44.73
MobileNet-v2 [27]	89.98	77.67	70.59	71.37	93.10	86.57	79.88	80.67	75.80	70.93	57.78	59.10	74.58	67.57	56.02	57.44
EfficientNet-D7 [30]	90.02	78.31	72.64	73.46	92.79	86.93	79.51	80.98	76.27	71.98	59.16	60.11	75.77	67.21	57.46	58.40
ViT-L/16 [4]	90.67	81.13	76.79	77.75	92.57	87.77	81.67	82.60	77.15	72.04	60.12	61.44	77.39	67.42	59.25	61.90
MFA-ViT	**92.43**	**85.43**	**81.01**	**83.79**	**92.81**	**89.40**	**85.95**	**86.88**	**80.32**	**73.70**	**67.56**	**69.09**	**80.18**	**68.80**	**62.77**	**65.15**

Table 6. Computational costs and model parameters on competing backbone structures.

Model	**#Param (M)**	**FLOPs (G)**	**Backbone**
VGGNet-16	171.68	31.22	CNN-based
MobileNet-v2	13.93	0.95	CNN-based
EfficentNet-D7	87.19	21.83	CNN-based
ViT-L/16	314.23	119.54	Transformer-based
MFA-ViT	98.81	31.29	Transformer-based

lining its effectiveness in enhancing the capabilities of our model. Our experiments were constrained to $\mathbf{P}'_*$ sizes of 24 and 32 due to memory limitations in our GPU hardware. While our findings suggest that this size offers performance advantages, further exploration with different sizes is warranted. This highlights the significance of careful hyperparameter tuning in multimodal systems, especially when hardware constraints come into play.

8.3. Impacts of Network Backbones

In addition, this study investigates various models trained with $\mathbf{I}_f$ and $\mathbf{I}_p$ modalities using identical $\mathcal{L}_{\text{total}}$ loss function. The networks utilize several backbone architectures, encompassing both CNN-based and transformer-based models. CNN-based model include VGGNet-16 [29], MobileNet-v2 [27], and EfficientNet-B7 [30], while transformer-based models include ViT-L/16 [4] and our MFA-ViT without MPT.

Table 5 presents the results, highlighting that MFA-ViT, even without a prompt strategy, consistently outperforms other models in terms of rank-1 accuracy across all datasets. We observe that this advantage is linked to the MFA layer. Notably, despite the transformer-based architecture of ViT-L/16, MFA-ViT exhibits superior performance. This discrepancy can be attributed primarily to the comparative ineffectiveness of its attention layer in aggregating multimodal features compared to our approach.

In essence, this study underscores the argument that relying solely on contrastive loss function without integrating the customized architectural design of MFA-ViT. The results indicate that the contrastive loss function exhibits limitations in grappling with intricate aspects of sophisticated tasks, such as feature fusion, alignment, and optimizing mutual information between modalities, which are paramount for achieving enhanced intra- and cross-modality learning outcomes.

In Table 6, we also present a comprehensive overview of each model's characteristics, including their total parameter size (*#Param.*) and floating-point operations per second (FLOPs). Notably, models like MobileNet-v2 and EfficientNet-D7 demonstrate a favorable reduction in parameter sizes and lower computational demands for FLOPs compared to our MFA-ViT. Despite these advantages in computational efficiency, their performance consistently lags behind that of MFA-ViT. This underscores that our model delivers high performance and offers practicality and scalability, making it well-suited for real-world applications.