Trading-off Mutual Information on Feature Aggregation for Face Recognition

Mohammad Akyash, Ali Zafari, Nasser M. Nasrabadi
Deptartment of Computer Science & Electrical Engineering, West Virginia University, WV USA
{ma00098,az00004}@mix.wvu.edu,{nasser.nasrabadi}@mail.wvu.edu

***Abstract*—Despite the advances in the field of Face Recognition (FR), the precision of these methods is not yet sufficient. To improve the FR performance, this paper proposes a technique to aggregate the outputs of two state-of-the-art (SOTA) deep FR models, namely ArcFace and AdaFace. In our approach, we leverage the transformer attention mechanism to exploit the relationship between different parts of two feature maps. By doing so, we aim to enhance the overall discriminative power of the FR system. One of the challenges in feature aggregation is the effective modeling of both local and global dependencies. Conventional transformers are known for their ability to capture long-range dependencies, but they often struggle with modeling local dependencies accurately. To address this limitation, we augment the self-attention mechanism to capture both local and global dependencies effectively. This allows our model to take advantage of the overlapping receptive fields present in corresponding locations of the feature maps. However, fusing two feature maps from different FR models might introduce redundancies to the face embedding. Since these models often share identical backbone architectures, the resulting feature maps may contain overlapping information, which can mislead the training process. To overcome this problem, we leverage the principle of Information Bottleneck to obtain a maximally informative facial representation. This ensures that the aggregated features retain the most relevant and discriminative information while minimizing redundant or misleading details. To evaluate the effectiveness of our proposed method, we conducted experiments on popular benchmarks and compared our results with state-of-the-art algorithms. The consistent improvement we observed in these benchmarks demonstrates the efficacy of our approach in enhancing FR performance. Moreover, our model aggregation framework offers a novel perspective on model fusion and establishes a powerful paradigm for feature aggregation using transformer-based attention mechanisms.**

***Index Terms*—Face recognition, Feature aggregation, Transformer, Cross-attention, Information bottleneck technique**

I. Introduction

The increased attention towards Face Recognition (FR) algorithms [1], [2], [3], [4], [5] in recent years can be attributed to several factors. One of the primary catalysts has been the rising demand for reliable and efficient face recognition systems in various domains, including security [6], surveillance [7], and identity verification [8]. As a result, researchers and practitioners alike have been actively exploring ways to enhance FR algorithms to meet these evolving needs.

Large-scale datasets have played a pivotal role in driving advancements in FR [1]. These datasets comprise vast collections of annotated face images, often containing millions of samples from diverse sources. The availability of such comprehensive data allows researchers to train FR algorithms on a rich variety of facial features, appearances, and scenarios. By leveraging these datasets, FR algorithms can learn to generalize better and exhibit improved performance when faced with real-world challenges, such as variations in lighting conditions, poses, expressions, and occlusions. In addition to large-scale datasets, novel loss functions have been instrumental in improving FR performance [9]. Loss functions define the objective that FR algorithms aim to optimize during training. Traditional loss functions, such as the softmax loss, have been enhanced or replaced with more sophisticated alternatives. For instance, the triplet loss [10] and its variants facilitate the learning of discriminative feature representations by encouraging closer proximity for images of the same identity and pushing images of different identities further apart in the embedding space. Other loss functions, such as center loss [11], focus on minimizing the intra-class variations while emphasizing inter-class separability. Most current FR methods (e.g., SphereFace [3], CosFace [2], and ArcFace [1]) focus on applying a margin penalty to the Softmax loss function to allow the network to extract more discriminative features. Recently, AdaFace [4] proposed a new loss function that considers image quality during the training process and emphasizes on recognizable low quality and high quality images.

Moreover, the development of new network architectures has significantly contributed to the progress in FR performance. Convolutional neural networks (CNNs) have revolutionized FR by effectively capturing facial features and patterns through hierarchical layers. Researchers have proposed various architectures, such as VGGNet [12], ResNet [13], InceptionNet [14], and more recently, efficient models like MobileNet [15] and EfficientNet [16], each designed to extract increasingly informative representations from face images. It is worth mentioning that advancements in hardware have also played a crucial role in facilitating the progress of FR algorithms. The availability of powerful GPUs, TPUs, and other specialized hardware accelerators has enabled researchers to train larger and more complex models efficiently. This computational power has expedited the experimentation process and allowed for more extensive exploration of network architectures, hyperparameters, and training techniques. Consequently, FR algorithms have benefited from faster training times, accelerated inference speeds, and the ability to handle

1946-0759/23/$31.00 ©2023 IEEE
DOI 10.1109/ICMLA58977.2023.00309

large-scale datasets effectively.

Despite the progress of in the field of FR, performance of the models is still not satisfactory. *Deep ensemble models* [17] mix the outputs of several independently trained methods to improve the generalization capability of the overall combination. Such ensemble models may significantly increase the accuracy of a single classifier in predicting unknown samples with high flexibility. In this paper, we exploit the transformer attention mechanism to fuse two identical networks trained with ArcFace [1] and AdaFace [4] loss functions.

Transformers, originally introduced for machine translation, have demonstrated exceptional performance across a wide range of natural language processing (NLP) tasks [18]. However, their potential extends beyond NLP, as exemplified by the Vision Transformer (ViT) [19], which utilizes self-attention mechanisms for image recognition. ViTs have gained significant popularity and have been successfully applied to various computer vision tasks such as image classification [20], compression [21], object detection [22], and video processing [23]. In the field of computer vision [24], feature fusion plays a pivotal role in enhancing model accuracy by combining information from multiple sensors and modalities, resulting in more robust and comprehensive analysis [25]. While transformers have been employed for feature fusion in tasks like image processing [26], [27], it is important to recognize that their main weakness lies in their inherent focus on modeling long-range dependencies between different components of visual data [28]. However, when it comes to feature aggregation, preserving local dependencies becomes vital, as there is often a shared receptive field between corresponding regions in feature maps. To address this limitation and enable effective feature aggregation, we propose a network that combines self-attention and cross-attention techniques. By leveraging these mechanisms, our network can aggregate feature maps both globally and locally. This approach allows us to effectively capture and combine both the local and global interactions between two feature maps, taking into account the specific characteristics and dependencies of the visual data. By incorporating self-attention and cross-attention techniques, our proposed network enhances the feature aggregation process by explicitly considering local dependencies, which are crucial for accurate representation learning. This enables our method to achieve more comprehensive and informative feature representations, leading to improved performance in various computer vision tasks, including face recognition.

The Information Bottleneck (IB) principle, introduced by [29], highlights the trade-off between learning a compact representation and achieving satisfactory prediction performance [30]. When combining two deep models, there is a risk of introducing redundancy into the ensemble model's output, potentially misleading the training process. To address this concern, we leverage the IB principle to obtain a compressed yet informative aggregated representation of the two model features. To achieve this, we incorporate a regularization term into the loss function that suppresses irrelevant information in the aggregated representation. This regularization encourages the network to focus on capturing essential and discriminative features while disregarding redundant or irrelevant details. By explicitly incorporating the IB principle into our model, we aim to strike a balance between compactness and performance, resulting in a more effective and efficient representation. The original IB method involves computationally expensive calculations of mutual information between the input, latent representation, and output. To mitigate this challenge, we adopt the concept of Variational Information Bottleneck (VIB) [31]. The VIB approach approximates the mutual information by introducing a variational lower bound, enabling more efficient computation and scalability to large-scale datasets. We evaluate our proposed model on a range of benchmark datasets, including AgeDB [32], CFP-FP [33], CPLFW [34], CALFW [35], LFW [36], IJB-B [37], and IJB-C [38]. Through these evaluations, we demonstrate significant improvements in performance compared to state-of-the-art (SOTA) algorithms, validating the effectiveness of our approach. By leveraging the IB principle and adopting the VIB framework, we address the challenge of redundancy in ensemble models while achieving a compressed and informative aggregated representation. The experimental results across various datasets underscore the superiority of our method, highlighting its potential for advancing the field of face recognition and outperforming existing state-of-the-art techniques.

To sum up, the contributions of this work are as follows:

1) A novel local-global transformer-based neural network is proposed to aggregate the output features of the ArcFace and AdaFace methods.
2) By employing the information bottleneck principle, we declare that the final output feature embedding is refined and does not have redundancies. In other words, our loss function guides the fusion network to suppress irrelevant information in the representation.
3) To demonstrate the efficacy of the proposed method, we perform extensive experiments on publicly-available datasets. Results confirm our technique performs well across various benchmarks.

The remainder of this paper is as follows: In Section III, we elaborate on our method. In Section IV, we evaluate our model and compare it to the SOTA methods. Finally, in Section V we conclude the paper.

II. Related Works

A. Face Recognition methods

In face recognition, a common margin-based loss functions aim to improve the discriminative power of the learned features by explicitly enforcing a margin between different classes in the feature space. It is commonly used in face recognition tasks to enhance the inter-class separability. In this subsection we introduce the recent advances in margin-based loss functions.

SphereFace [3] introduces a novel angle-based softmax loss that incorporates a margin function to enhance the discriminative power of the learned features. The margin function is designed to ensure that the features belonging to different

classes are well separated in the angular space. The angle-based softmax loss used in SphereFace can be defined as:

$$L_{\text{sphere}} = -\log\left(\frac{\exp(s(\cos(\theta_{yi} - m)))}{\exp(s(\cos(\theta_{yi} - m))) + \sum_{j \neq yi}\exp(s\cos(\theta_j))}\right), \tag{1}$$

where L_{sphere} is the loss function, s is a scaling factor, θ_{yi_m} is the angle between the input feature and the weight vector of the ground truth class yi after applying a margin m, and θ_j is the angle between the input feature and the weight vector of class j. The margin m is introduced to increase the angular separation between different classes. CosFace [2] focuses on enhancing the margin-based loss by incorporating the cosine similarity metric. The margin function used in CosFace is designed to increase the angular separation between different classes in the feature space. The CosFace loss function can be defined as:

$$L_{\text{cos}} = -\log\left(\frac{\exp(s(\cos(\theta_{yi}) - m)))}{\exp(s(\cos(\theta_{yi}) - m))) + \sum_{j \neq yi}\exp(s\cos(\theta_j))}\right), \tag{2}$$

where L_{cos} is the loss function, s is a scaling factor, θ_{yi} is the angle between the input feature and the weight vector of the ground truth class yi, and θ_j is the angle between the input feature and the weight vector of class j. The margin m is added to increase the angular margin between classes. ArcFace [1] builds upon the concept of using a margin function within the softmax loss to improve the discriminative capacity of the learned features. The margin function in ArcFace is designed to enforce large angular separations between classes in the feature space. The ArcFace loss function can be defined as:

$$L_{\text{arc}} = -\log\left(\frac{\exp(s(\cos(\theta_{yi} + m)))}{\exp(s(\cos(\theta_{yi} + m))) + \sum_{j \neq yi}\exp(s\cos(\theta_j))}\right), \tag{3}$$

where L_{arc} is the loss function, s is a scaling factor, θ_{yi} is the angle between the input feature and the weight vector of the ground truth class yi, and θ_j is the angle between the input feature and the weight vector of class j. The margin m is added to increase the angular separation between classes.

AdaFace [4] introduces an adaptive margin-based approach that dynamically adjusts the margin for each training sample, leading to improved discriminability. The margin adaptation process in AdaFace aims to handle intra-class variations by assigning larger margins to challenging samples and smaller margins to easier ones. The AdaFace loss function can be defined as:

$$L_{\text{ada}} = -\log\left(\frac{\exp(s(\cos(\theta_{yi} + m_i)))}{\exp(s(\cos(\theta_{yi} + m_i))) + \sum_{j \neq yi}\exp(s\cos(\theta_j))}\right), \tag{4}$$

Where L_{ada} is the loss function, s is a scaling factor, θ_{yi} is the angle between the input feature and the weight vector of the ground truth class yi, θ_j is the angle between the input feature and the weight vector of class j, and m_i is the dynamically adjusted margin for each training sample. The margin m_i is computed based on the difficulty or intra-class variations of the sample. By employing margin-based loss functions, including the specific forms used in SphereFace, CosFace, ArcFace, and the adaptive margin approach in AdaFace, these techniques aim to enhance the discriminative power of face recognition models and improve their performance in challenging scenarios.

One of the key advantages of these methods is their ability to enhance the discriminative power of the learned features. By incorporating margin functions within the softmax loss, these techniques effectively increase the angular separations between different classes in the feature space. This leads to better discrimination between individuals, resulting in more accurate face recognition. Another advantage of these methods is their robustness to variations commonly encountered in face recognition, such as pose variations, lighting conditions, and occlusions. By incorporating margin functions and angular constraints, these techniques encourage the learned features to be less affected by variations, resulting in improved robustness and generalization capabilities.

B. Ensemble Learning methods

In machine learning, an ensemble model is a technique that combines multiple individual models to make predictions. The idea behind ensemble models is that the combination of several weak models can result in a stronger and more accurate predictor. In this subsection, we review some of the well-known ensemble learning methods. Random Forest is a popular ensemble method introduced by [39]. It combines multiple decision trees, where each tree is trained on a random subset of the data and features. Random Forest has been widely used in various domains due to its robustness and ability to handle high-dimensional data. Gradient Boosting Machines (GBM) [40] is another powerful ensemble technique that builds an ensemble of weak prediction models, typically decision trees, in a sequential manner. Each model is trained to correct the mistakes made by the previous models. Notable GBM implementations include XGBoost [41] and LightGBM [42], which have gained significant popularity due to their efficiency and performance. Deep ensembles, as discussed previously, combine multiple deep learning models to form an ensemble. These models are typically deep neural networks trained independently and combined using techniques such as averaging or voting. Deep ensembles have been applied to various domains, including image classification, natural language processing, and reinforcement learning. Bayesian Model Averaging (BMA) [43] is a probabilistic ensemble approach that assigns weights to individual models based on their performance on a validation set. Bayesian techniques are used to estimate the weights, and the final prediction is obtained by averaging the predictions of all models weighted by their probabilities. Stochastic Weight Averaging (SWA) is a recent ensemble technique proposed by [44]. It involves averaging the weights of multiple models during training rather

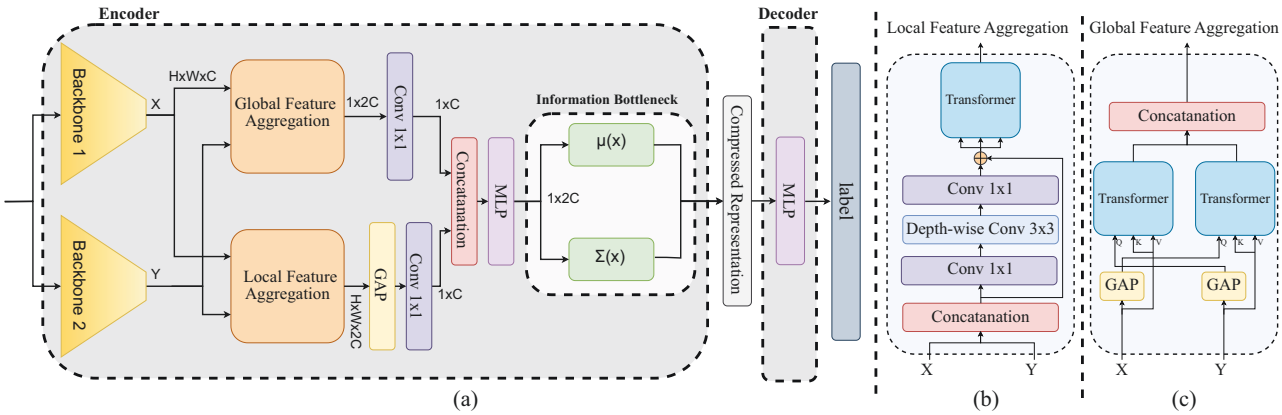

Fig. 1: (a) The overall architecture of the proposed method. The input image is fed to two pre-trained FR backbones to obtain two feature maps $X, Y \in R^{H\times W\times C}$. Next, the feature maps are aggregated with the introduced global and local transformer-based modules and then, we exploit the IB technique to obtain a compressed representation of the image. (b) The local feature aggregation module. (c) The global aggregation module.

than their predictions. This method has been shown to improve generalization and robustness of deep learning models.

III. Method

Figure 1(a) illustrates the comprehensive architecture of our proposed method. To begin, we extract feature maps from the final convolutional layer of each pre-trained face recognition (FR) backbone. These feature maps serve as input to two parallel modules within our approach, enabling simultaneous local and global information aggregation. To facilitate this aggregation process, we employ a transformer encoder architecture as depicted in Figure 2. The transformer encoder acts as the fundamental building block of our feature aggregation modules, allowing for efficient capturing and integration of local and global facial information. Following the feature aggregation step, we concatenate the outputs from the local and global modules and leverage the Information Bottleneck (IB) technique. This technique enables us to achieve a compressed representation that retains the essential discriminative information while removing redundancies, optimizing the overall performance of the face recognition model. Finally, we decode the compressed representation to obtain the corresponding labels, allowing for accurate classification. During the inference phase, the compressed representation is utilized for 1:1 face verification, enabling efficient and reliable matching between pairs of face images. By adopting this comprehensive architecture, our method effectively combines local and global information, leverages the power of transformers for feature aggregation, incorporates the benefits of the Information Bottleneck principle for compression, and ultimately enables accurate face recognition and verification tasks.

A. *Information Bottleneck method*

In our loss function, we have used the IB principle [29] to achieve a compressed and informative fused representation of the images. In a classification task, we need to learn a representation that is maximally compressed with regard to the input and maximally informative about the output. The IB principle is defined below:

$$\mathcal{L}_{IB}(\theta) = \beta I(\hat{X}, X; \theta) - I(\hat{X}, Y; \theta), \tag{5}$$

where $I(.,.)$ denotes the mutual information, and X, $\hat{X}$, and Y represent the input, bottleneck representation, and corresponding labels, respectively.

*1) **Variational Information Bottleneck**:* The main drawback of the IB principle is that the computation of mutual information is cumbersome, especially for continuous and high-dimensional variables. Recently, remarkable improvements have allowed the computation of MI in an efficient manner [31], [45]. In [31] a variational bound is presented to approximate the IB objective. This bound is defined below:

$$\begin{aligned}\mathcal{L}_{IB}(\theta) &= \beta I(\hat{X}, X; \theta) - I(\hat{X}, Y; \theta) \\ &\leq \beta \int p(x)p_\theta(\hat{x}|x)log\frac{p_\theta(\hat{x}|x)}{r(\hat{x})}dxd\hat{x} \\ &\quad - \int p(x)p(y|x)p_\theta(x|\hat{x})log_{q_\phi}(y|\hat{x})dxdyd\hat{x}\,,\end{aligned} \tag{6}$$

where $p_\theta(\hat{x}|x)$ is the estimation for the posterior probability, $r(\hat{x})$ is a normal distribution and $q_\theta(y|\hat{x})$ is the estimation of distribution Y. The loss function is then defined below:

$$\begin{aligned}\mathcal{L}_{VIB}(\theta, \phi) &= \beta\mathbb{E}_x[KL(p_\theta(\hat{x}|x), r(\hat{x}))] \\ &\quad + \mathbb{E}_{\hat{x}\sim p_\theta(\hat{x}|x)}[-log(q_\phi(y|\hat{x}))]\,,\end{aligned} \tag{7}$$

where $D_{KL}(.,.)$ denotes the *Kullback-Leibler divergence*.

B. *Attention-Based Fusion Architecture*

1) ***Local Feature Aggregation****:* While a transformer attention mechanism excels at capturing long-range dependencies across input tokens, it does not inherently emphasize the interaction between tokens within a local region. In our feature fusion approach, where we compute feature maps using two identical backbones, the pixels within a local region of the feature maps share common receptive fields in the original image. Therefore, it is crucial to design a model that can effectively capture the relationships between elements in corresponding local positions of the two feature maps.

To address this requirement, we propose a variation of the transformer architecture that specifically focuses on modeling the local context between different parts of the input features. Figure 1(b) illustrates the architecture of our local feature aggregation module. Inspired by the work of [28], we modify the conventional transformer to enhance its capability to capture local context. To enable the transformer to effectively capture local context, we begin by concatenating the two feature maps, denoted as X and Y, resulting in a composite feature map F of shape $H \times W \times 2C$. We then apply a sequence of operations, including a $Conv(1\times1)$, a depth-wise convolution, and another $Conv(1\times1)$, followed by adding F to the output. This series of operations is designed to establish local context fusion within the transformer, a capability that the conventional transformer lacks. Subsequently, the resulting feature map is flattened to F' of shape $HW \times C$ and fed into the transformer module. In this configuration, the number of tokens is HW, each having a size of C. By incorporating the depth-wise convolution, we aim to effectively capture and integrate local context within the transformer, enabling it to model the relationships between elements in local regions. It is important to note that the introduction of the depth-wise convolution does not significantly increase the computational complexity compared to the original transformer. This is due to the low computational overhead associated with depth-wise convolutions. Therefore, our modified architecture remains computationally efficient while effectively capturing local context and enhancing feature fusion. By leveraging the local feature aggregation module, we enable our model to capture both global and local interactions within the feature maps. This comprehensive understanding of relationships between elements contributes to more accurate and robust feature fusion, ultimately enhancing the performance of our approach in face recognition.

2) ***Global Feature Aggregation****:* While Transformers are known for their ability to model global interactions among tokens, they typically consider the relationship of a single token with other tokens at a time, neglecting the potential for fully utilizing the interaction among all elements within a feature. To enhance performance and enable a more comprehensive understanding of the global relations between elements in two features, we modify the Transformer network accordingly.

The global feature fusion module, depicted in Figure 1(c), plays a crucial role in this enhancement. The input to this module consists of two feature maps, denoted as X and Y, both of shape $H \times W \times C$. To facilitate interaction between the feature maps, we apply Global Average Pooling (GAP) to each feature map, resulting in two feature vectors, G_X and G_Y, both of shape $1 \times C$. To enable the feature maps to interact with each other, we utilize the feature vector G_Y as the query of one Transformer, and G_X as the query of another Transformer. The keys and values for the Transformers are derived from Y and X, respectively. The result of applying a $Conv(1 \times 1)$ operation to $GAP(Y)$, X, and Y produces K and V, both of shape $HW \times C$. The multi-head attention mechanism is then applied, resulting in Z_1, an output of shape $1 \times C$. Following the attention step, the output passes through the "Add and Norm," "FeedForward," and "Add and Norm" layers. These operations further refine the features and contribute to the final representations, resulting in two feature vectors, F_1 and F_2, both of shape $1 \times C$. The two attended feature vectors, F_1 and F_2, are concatenated to form a resulting feature vector, denoted as F, of shape $2\times C$. This fusion allows for the integration of the attended information from both feature maps, capturing a richer representation of the global relations between the elements. Importantly, in this fusion process, the query of each transformer module is the average of all the pixels in a channel, enabling a global perspective for attention. Consequently, the values are weighted based on the interaction between the "globally designed" query and the key. This attention mechanism ensures that the interaction among the elements is effectively captured and utilized to enhance the overall representation.

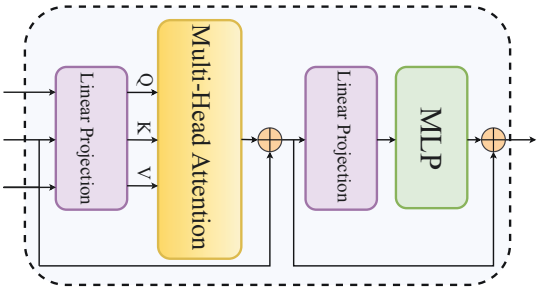

Fig. 2: The transformer encoder architecture, the base block of our proposed local and global feature aggregation modules.

C. *Objective Function*

The VIB objective function serves as the loss function for training our network. In the literature of variational autoencoders, the encoder and decoder are denoted as $p_\theta(\hat{x}|x)$ and $q_\phi(y|\hat{x})$ respectively. To align with this convention, we rewrite Equation 7 as follows:

$$\mathcal{L}_{VIB}(\theta,\phi) = \beta\mathcal{L}_{encoder}(\theta) + \mathcal{L}_{decoder}(\phi). \tag{8}$$

In our network architecture, the backbones and the global-local feature aggregation components collectively function as our encoder. On the other hand, the decoder consists of a

fully connected layer that connects the bottleneck representation $\hat{x}$ with the corresponding labels. Within Equation 8, the $\mathcal{L}_{decoder}$ term represents the cross-entropy loss, which measures the dissimilarity between the predicted labels and the true labels. On the other hand, the $\mathcal{L}_{encoder}$ term acts as a regularization term, imposing constraints on the network to encourage the removal of redundant information from the learned representation $\hat{x}$. This regularization facilitates a more compact and informative representation. During the training process, the backbones remain fixed, so the trainable weights θ correspond to the fusion network. Additionally, the weights ϕ represent those of the fully connected layer. The hyperparameter β determines the extent of compression applied to the learned representation. To approximate the parameters of $p_\theta(z|x)$, we make use of the approximations $\mu(x)$ and $\Sigma(x)$. During training, the compressed representation $\hat{x}$ is sampled from $p_\theta(z|x)$, while during inference, we utilize $\mu(x)$ as the representation for the input image. By formulating our loss function in this manner and incorporating the VIB objective, we enable the network to simultaneously optimize the decoder for accurate classification and the encoder for efficient representation compression. This framework allows us to achieve a balance between preserving predictive information and eliminating redundancies, ultimately enhancing the face recognition performance of our model.

IV. Experimental Setting and Results

A. Datasets

We have incorporated a segment of WebFace12M [50], comprising over 5 million facial images, into our training dataset. During the testing phase, we evaluated our model using diverse datasets with varying image qualities to assess its robustness and generalization capabilities. For the high-quality image datasets, we utilized AgeDB [32], CFP-FP [33], CPLFW [34], CALFW [35], LFW [36], IJB-B [37], and IJB-C [38] datasets, which are widely recognized benchmarks within the face recognition (FR) community. These datasets are known for their well-captured and well-aligned images, providing an ideal evaluation environment for assessing the performance of FR methods. To ensure consistent and standardized evaluation, we pre-processed each dataset using the techniques outlined in [51], which includes face detection and alignment procedures. Furthermore, we followed the settings defined in [1] to perform rescaling and alignment, ensuring fair and comparable evaluation conditions across different methods. To align with the evaluation practices of state-of-the-art methods, we reported the 1:1 verification accuracy for the aforementioned datasets. Additionally, we presented the True Accept Rate (TAR) at a False Accept Rate (FAR) of $1e-4$, which provides a comprehensive measure of the model's performance in distinguishing genuine matches from impostor matches. By adhering to established evaluation standards and employing a diverse range of datasets, including those with both high-quality and low-quality images such as IJB-B and IJB-C, we aim to demonstrate the effectiveness and versatility of our proposed method across various real-world scenarios.

B. Implementation details

For the backbones, we exploit the ResNet100, pre-trained with ArcFace [1] and AdaFace [4] losses with the same training dataset as ours. Using stochastic gradient descent (SGD), the entire network is trained for 24 epochs with the $\mathcal{L}_{VIB}$ loss function. The learning rate begins at 0.1 and is decreased by a factor of 10 at 10^{th}, 16^{th}, and 22^{th} epochs. For the training phase, each image and its corresponding label (as a one-hot vector) are fed to the network. During the inference phase, the pair is given to the network, and the cosine distance is then computed between the representations as a metric. For experiment, first, we evaluate only the global branch then we use the local branch, and at the end, we take advantage of both the local and global branches. In the last experiment, $\mathcal{L}_{VIB}$ is used, and the compressed representation length is $K = 512$.

C. Comparison with the SOTA methods

In Table I, we exhibit the performance of our algorithm compared to the state-of-the-art techniques, and as we can see in the table, our method outperforms these algorithms. We evaluate our model with global, local, and both local and global modules, and we achieved the best results when we used both modules simultaneously. To demonstrate the effectiveness of the loss function $\mathcal{L}_{VIB}$, we conduct experiments with and without VIB, and the results indicate that by using this loss function, we can achieve better performance. Furthermore, in the next part, we explain how tuning the hyper-parameter β allows us to achieve the optimum performance on both high and mixed-quality datasets. The hyper-parameter β controls the trade-off between the reconstruction fidelity and the disentanglement of the latent representations. Through careful tuning of β, we are able to strike a balance that ensures high performance across various datasets.

D. Effect of hyper-parameter β on accuracy

We examine how the hyper-parameter β in $\mathcal{L}_{VIB}$ affects the face recognition performance by controlling the trade-off between preserving predictive information and compression in the latent representation. A low β value indicates a greater emphasis on preserving predictive information, while a higher β value prioritizes compression and eliminates redundancies. To investigate the impact of β on performance, we trained our model with various values of this parameter. Fig. 3 illustrates the verification accuracy on the validation datasets, providing insights into the relationship between β and performance. Our experiments reveal that different datasets exhibit varying sensitivity to β due to differences in image quality. For high-quality datasets such as AgeDB and CFP-FP, a higher β value is required. These datasets contain rich contextual information, necessitating a greater degree of compression in their representations. By increasing β, we can effectively eliminate redundancies and achieve improved performance on these high-quality images. Conversely, for lower-quality images, a lower β value proves more effective. These images contain less contextual information and may benefit from a more informative representation that preserves a higher degree

TABLE I: The results of our proposed method are compared to the SOTA methods for the 1:1 face verification task.

Method	High Quality					Mixed Quality	
	LFW	CFP-FP	CPLFW	AgeDB	CALFW	IJB-B	IJB-C
ArcFace [1]	99.83	98.27	92.08	98.28	95.45	94.25	96.03
AdaFace [4]	99.82	98.49	93.53	98.05	96.08	95.67	96.89
CurricularFace [46]	99.80	98.37	93.13	98.32	96.20	94.80	96.10
MagFace [47]	99.83	98.46	92.87	98.17	96.15	94.51	95.97
BroadFace [48]	**99.85**	98.63	93.17	98.38	96.20	94.97	96.37
SCF-ArcFace [49]	99.82	98.40	93.16	98.30	96.12	94.74	96.09
Fusion + Global	99.83	98.54	93.46	98.37	96.15	95.54	96.75
Fusion + Local	99.83	98.46	93.50	98.25	96.20	95.63	96.84
Fusion + Global + Local	99.83	98.50	93.53	98.30	96.18	95.65	96.89
Fusion + Global + Local + VIB	**99.85**	**98.87**	**93.78**	**98.60**	**96.25**	**95.83**	**97.11**

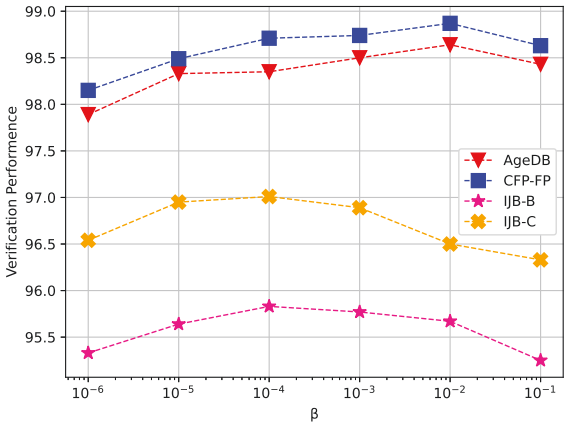

Fig. 3: The verification performance versus the hyper-parameter β. To achieve optimal performance for high quality images we need higher β (more compression) due to the higher amount of contextual information.

of target-related details. By reducing β, we strike a balance that allows for more preservation of relevant information in the face recognition process for lower-quality datasets. In summary, our investigation demonstrates the importance of tuning β in $\mathcal{L}_{VIB}$ to achieve optimal face recognition performance. The appropriate choice of β depends on the dataset characteristics, with higher values suitable for high-quality images and lower values preferred for lower-quality images. This flexibility enables our model to adapt to different datasets and maximize performance across a range of image qualities.

V. Conclusion

In this paper, we propose a transformer-based architecture to enhance face recognition performance by aggregating the output features of two pre-trained networks. Transformers have limitations in capturing local interactions, so we divided the fusion module into local and global feature aggregation components. To address potential redundancies in the aggregated features, we leverage the Information Bottleneck (IB) principle to achieve a maximally informative and compressed representation. We evaluate our model on various benchmarks and demonstrate its superiority over SOTA methods. Overall, our approach effectively addresses the limitations of transformers in capturing local interactions by dividing the fusion module and leveraging the IB principle. The experimental results showcase the improved performance of our model compared to existing methods. This paper contributes to the field of face recognition by proposing an enhanced architecture for real-world scenarios.

VI. Acknowledgment

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2022-21102100001. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

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

[2] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *CVPR*, 2018, pp. 5265–5274.

[3] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *CVPR*, 2017, pp. 212–220.

[4] M. Kim, A. K. Jain, and X. Liu, "Adaface: Quality adaptive margin for face recognition," in *CVPR*, 2022, pp. 18 750–18 759.

[5] N. A. Talemi, H. Kashiani, S. R. Malakshan, M. S. E. Saadabadi, N. Najafzadeh, M. Akyash, and N. M. Nasrabadi, "Aaface: Attribute-aware attentional network for face recognition," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 1940–1944.

[6] M. Rakhra, D. Singh, A. Singh, K. D. Garg, and D. Gupta, "Face recognition with smart security system," in *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 2022, pp. 1–6.

[7] M. Karpagam, R. B. Jeyavathana, S. K. Chinnappan, K. Kanimozhi, and M. Sambath, "A novel face recognition model for fighting against human trafficking in surveillance videos and rescuing victims," *Soft Computing*, pp. 1–16, 2022.

[8] B. Adami, S. Tehranipoor, N. M. Nasrabadi, and N. Karimian, "A universal anti-spoofing approach for contactless fingerprint biometric systems," in *2023 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2023, pp. 1–8.

[9] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," *arXiv preprint arXiv:1612.02295*, 2016.

[10] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*. Springer, 2015, pp. 84–92.

[11] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*. Springer, 2016, pp. 499–515.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[15] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[16] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[17] M. A. Ganaie, M. Hu, A. Malik, M. Tanveer, and P. Suganthan, "Ensemble deep learning: A review," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105151, 2022.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[20] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *ICCV*, 2021, pp. 357–366.

[21] A. Zafari, A. Khoshkhahtinat, P. Mehta, M. S. E. Saadabadi, M. Akyash, and N. M. Nasrabadi, "Frequency disentangled features in neural image compression," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 2815–2819.

[22] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in *ICCV*, 2021, pp. 11 936–11 945.

[23] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," in *ICCV*, 2021, pp. 3163–3172.

[24] M. Babic, M. A. Farahani, and T. Wuest, "Image based quality inspection in smart manufacturing systems: A literature review," *Procedia CIRP*, vol. 103, pp. 262–267, 2021.

[25] M. Alipour, I. La Puma, J. Picotte, K. Shamsaei, E. Rowell, A. Watts, B. Kosovic, H. Ebrahimian, and E. Taciroglu, "A multimodal data fusion and deep learning framework for large-scale wildfire surface fuel mapping," *Fire*, vol. 6, no. 2, p. 36, 2023.

[26] W. Zhang, F. Qiu, S. Wang, H. Zeng, Z. Zhang, R. An, B. Ma, and Y. Ding, "Transformer-based multimodal information fusion for facial expression analysis," in *CVPR*, 2022, pp. 2428–2437.

[27] L. Zhou and Y. Luo, "Deep features fusion with mutual attention transformer for skin lesion diagnosis," in *ICIP*. IEEE, 2021, pp. 3797–3801.

[28] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, "Incorporating convolution designs into visual transformers," in *ICCV*, 2021, pp. 579–588.

[29] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[30] S. Mohamadi, G. Doretto, and D. A. Adjeroh, "More synergy, less redundancy: Exploiting joint mutual information for self-supervised learning," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 1390–1394.

[31] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[32] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "Agedb: the first manually collected, in-the-wild age database," in *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 51–59.

[33] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–9.

[34] T. Zheng and W. Deng, "Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments," *Beijing University of Posts and Telecommunications, Tech. Rep*, vol. 5, no. 7, 2018.

[35] T. Zheng, W. Deng, and J. Hu, "Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments," *arXiv preprint arXiv:1708.08197*, 2017.

[36] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[37] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen *et al.*, "Iarpa janus benchmark-b face dataset," in *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 90–98.

[38] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney *et al.*, "Iarpa janus benchmark-c: Face dataset and protocol," in *2018 international conference on biometrics (ICB)*. IEEE, 2018, pp. 158–165.

[39] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[40] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[41] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[42] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[43] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and ei george, and a rejoinder by the authors," *Statistical science*, vol. 14, no. 4, pp. 382–417, 1999.

[44] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," *arXiv preprint arXiv:1803.05407*, 2018.

[45] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *ICML*. PMLR, 2018, pp. 531–540.

[46] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "Curricularface: adaptive curriculum learning loss for deep face recognition," in *CVPR*, 2020, pp. 5901–5910.

[47] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *CVPR*, 2021, pp. 14 225–14 234.

[48] Y. Kim, W. Park, and J. Shin, "Broadface: Looking at tens of thousands of people at once for face recognition," in *ECCV*. Springer, 2020, pp. 536–552.

[49] S. Li, J. Xu, X. Xu, P. Shen, S. Li, and B. Hooi, "Spherical confidence learning for face recognition," in *CVPR*, 2021, pp. 15 629–15 637.

[50] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du *et al.*, "Webface260m: A benchmark unveiling the power of million-scale deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 492–10 502.

[51] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *CVPR*, 2020, pp. 5203–5212.