Improving Face Recognition from Caption Supervision with Multi-Granular Contextual Feature Aggregation

Anonymous IJCB 2023 submission

Abstract

We introduce caption-guided face recognition (CGFR) as a new framework to improve the performance of commercial-off-the-shelf (COTS) face recognition (FR) systems. In contrast to combining soft biometrics (e.g., *facial marks, gender, and age) with face images, in this work, we use facial descriptions provided by face examiners as a piece of auxiliary information. However, due to the heterogeneity of the modalities, improving the performance by directly fusing the textual and facial features is very challenging, as both lie in different embedding spaces. In this paper, we propose a contextual feature aggregation module (CFAM) that addresses this issue by effectively exploiting the fine-grained word-region interaction and global image-caption association. Specifically, CFAM adopts a self-attention and a cross-attention scheme for improving the intra-modality and inter-modality relationship between the image and textual features. Additionally, we design a textual feature refinement module (TFRM) that refines the textual features of the pre-trained BERT encoder by updating the contextual embeddings. This module enhances the discriminative power of textual features with a cross-modal projection loss and realigns the word and caption embeddings with visual features by incorporating a visual-semantic alignment loss. We implemented the proposed CGFR framework on two face recognition models (ArcFace and AdaFace) and evaluated its performance on the Multimodal CelebA-HQ dataset. Our framework improves the performance of ArcFace from 16.75% to 66.83% on TPR@FPR=1e-4 in the 1:1 verification protocol.*

1. Introduction

Despite remarkable advancements in face recognition due to the adoption of margin-based loss functions [2, 14], face recognition in unconstrained scenarios remains a challenging problem [38]. The presence of covariate factors in an unconstrained environment, such as resolution, illumination, and pose, affects the face image quality, thus, decreasing the recognition performance. Providing auxiliary information, such as facial marks, gender, ethnicity, age, and skin color, to a face recognition (FR) system can improve its recognition performance [5, 39]. For example, in an unconstrained environment such as video surveillance, where a prevalent commercially-off-the-shelf (COTS) system performs poorly [38, 41], the application of soft biometrics has been proven to improve the performance of hard biometrics [30, 5].

Natural language captions that describe the visual contents of a face are an essential soft biometric trait for face recognition. In this study, we will explore whether we can boost the performance of a FR system using caption supervision. Our caption-guided face recognition (CGFR) has enormous potential in various applications, such as criminal and intelligence investigation, video surveillance, etc. For example, using a CGFR model, law-enforcement agencies can quickly retrieve the suspect face from a low-quality CCTV footage and a short description of eyewitnesses.

Although captions are rich, they face many challenges that limit their application in the biometric system. As natural language contains high-dimensional information, it is often much more abstract than images. A short textual description of a given face consisting of a few sentences is insufficient to describe all the minute details of the facial features. Consequently, CGFR is significantly different from other tasks such as cross-modal image-text retrieval (ITR) [13, 17] and image-text matching (ITM) [16], where the matching text has a description of the various objects, background scenes, styles, etc. Moreover, different people may have different captions for a particular face.

To improve the performance of the FR systems using CGFR, it is essential to find not only the semantic understanding of textual contents but also the proper association between visual and textual modalities. This is because the embedding space of images and text lies in different spaces due to the heterogeneity of the two modalities [17]. Aligning the image features with word embeddings is thus crucial, as it has a significant impact on the performance of a cross-modal fusion algorithm [17]. In this work, we fine-tune the state-of-the-art BERT model [4] to update the con-

textual associations among words in the caption by incorporating a visual-semantic alignment loss [36] and a cross-modal classification loss [40]. Finetuning the text encoder is essential because the BERT model was trained with different objectives than ours. Therefore, we finetune to achieve two objectives: (1) to learn visually aligned text embedding, i.e., to realign word and caption embeddings with visual information, and (2) to enhance the discriminative power of textual features.

However, a simple feature-level cross-modal fusion without fine-grained interaction between image-text tokens does not perform well. Therefore, we propose a novel module, namely, contextual feature aggregation module (CFAM), to effectively carry out the fine-grained word-region interaction and global image-caption association on two different granularities. There are mainly three networks in the proposed CFAM: caption-level context modeling, word-level context modeling, and a feature aggregation network. Both context modeling networks adopt a self-attention and a cross-attention mechanism. The self-attention mechanism increases the intra-modality relationship within each modality, while the cross-attention mechanism improves the inter-modality relationship between image and textual features. The inputs to the feature aggregation network are the context-enhanced features from the word and caption level context modeling.

We conduct our experiments on a benchmark text-to-face dataset, namely, Multi-Modal CelebA-HQ [33] (MM-CelebA). Sample image-caption pairs of the dataset are illustrated in Figure 1. In fact, the dataset is based on a subset of the CelebA dataset [21] that contains high-resolution images with very low variation. In our experiment, we remove the crucial face-alignment step and apply some preprocessing steps such as random sub-sampling, rotation, flip, etc., to augment our database as well as downgrade the image quality in order to mimic real-world low quality video surveillance scenarios. The verification rate of FR systems, such as ArcFace [2] and AdaFace [14], drops drastically on this preprocessed dataset because the images are corrupted with down-sampling and noise, which adversely affect their facial analysis procedure [41]. We then apply our CGFR framework to both systems. The experimental results demonstrate a remarkable performance leap over the COTS systems.

In this study our contributions are: (1) exploring a new paradigm to improve face recognition with natural language supervision, (2) proposing the CFAM module to exploit fine-grained interaction among local and global features using word and caption-level of granularity, (3) designing a textual feature refinement module (TFRM) to refine textual features and align them with visual content by fine-tuning the BERT encoder, and (4) conducting extensive experiments on the MMCelebA [33] dataset using the proposed

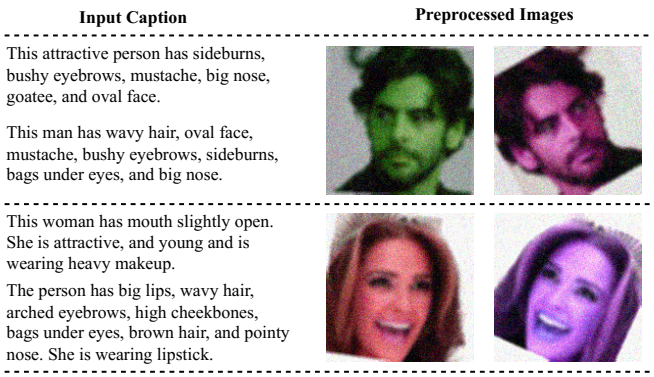

Figure 1. Sample image-caption pairs from the state-of-the-art Multi-Modal CelebA-HQ text-to-face dataset.

CGFR framework to demonstrate substantial improvements over existing FR systems. Finally, this work demonstrates excellent potential for caption-guided face recognition and provides a promising approach for further research.

The rest of this paper is organized as follows: an overview of the related works is presented in Section 2. A detailed description of the proposed method, including steps to finetune the BERT encoder, is presented in Section 3. In Section 4, we demonstrate the experimental evaluation of the CGFR framework. Finally, we summarize our results with some possible future research directions in Section 5.

2. Related Work

2.1. Soft Biometrics

Facial semantic attributes, such as facial marks, hair color, gender, ethnicity, age, and skin color, have been significantly exploited as auxiliary information to improve the performance of tasks such as face image retrieval and face verification. However, most of the prior works in the literature on improving face recognition using soft biometrics have been based on using categorical labels [5, 39]. Zhang *et al*. [39] integrated a set of five soft biometrics (ethnicity, gender, eyebrow, eye color, and hair color) with hard biometric systems. Compared to the baseline recognition rates at FAR = 0.001, their verification rate improved up to 15.5% when introducing all the soft biometrics and 16.4% when introducing gender information on the ugly part of the GBU database. Furthermore, authors in [5] empirically proved that a manual estimation of the six most discriminative soft biometric improves the relative performance of the FR systems (COTS Face++ and VGG-face) up to 40% over the LFW database.

2.2. Caption-Guided Face Recognition

Several early works have been proposed for caption-supervised face recognition [11, 7]. Huang *et al*. [11] improved state-of-the-art face recognition using web-scale im-

ages with captions by learning the feature space in an iterative label expansion process. However, they only employed captions to extrapolate the labels of the face identity.

Recently, with the development of generative adversarial networks (GANs) [6] and transformers [31], text-to-face synthesis [32, 29], and facial attribute editing [9, 33] with textual descriptions have gained increasing popularity. For example, TediGAN [33] uses latent code optimization of pre-trained StyleGAN for caption-guided facial image generation and manipulation. In contrast to these works, we introduce a new line of research by exploring the idea of using natural language captions to improve the performance of the FR systems. As there are no publicly available datasets that contain large-scale image-caption pairs for our task, we employ MMCelebA [33] dataset which has been widely used for text-to-face synthesis.

2.3. Attention Techniques

Recently, different attention mechanisms, such as self-attention, cross-attention, etc., have been extensively exploited in various multimodal tasks [18, 16, 36, 37]. Cross-attention or co-attention is an attention mechanism initially proposed in transformers [31] that interacts with two embedding sequences from different modalities (*e.g.*, text, image). Li *et al*. [18] propose a latent co-attention mechanism in which spatial attention relates each word to corresponding image regions. Also, Lee *et al*. [16] developed a stacked cross-attention network that learns the cross-modal alignments among all regions in an image and words in a sentence. Xu *et al*. [36] applied an attention mechanism to guide the generator to focus on different words while generating various image sub-regions. They also proposed a deep attentional multimodal similarity model (DAMSM) to improve the similarity between the generated images and the given descriptions. To re-weight the importance of local image regions in tasks such as image synthesis [36], image caption generation [35], image segmentation [28, 37], and image-text matching [16, 18], word-level attention has been employed. However, only employing word-level attention cannot ensure global semantic consistency due to the diversity of the text and image modalities. Global contextual information is also important as it provides more information on the visual content of the image, and context of the caption. Therefore, global image-caption attention should also be considered to drive the global features toward a semantically well-aligned context.

2.4. Multimodal Representation Learning

In recent years, dual-stream approaches, where the image and text encoder are trained on large-scale datasets individually with different cross-modality loss functions, have become widely popular in tackling various multimodal downstream tasks [25]. A lot of cross-modal loss functions such as contrastive [25, 19], triplet [16], word-region alignment [36], cross-modal projection [40], etc., have been proposed as part of the training objectives. Zhang *et al*. [40] proposed a novel projection loss that consists of two losses: a cross-modal projection matching (CMPM) loss for computing the similarity between image-text pairs and a cross-modal projection classification (CMPC) loss for learning a more discriminative visual-semantic embedding space. However, most of the dual-stream approaches in the literature cannot effectively and accurately exploit the fine-grained interaction among word-region features.

Furthermore, other researchers [37] has used image and textual features, extracted from separate encoders, which are often concatenated to be fed into a fusion module to learn joint representations. However, a simple fusion scheme may not be effective since the unaligned visual and word tokens lack prior relationships. Therefore, cross-modal interaction from local and global contexts is essential to improve multimodal fusion performance. For example, Niu *et al*. [24] map phrases-region and image-caption into a joint embedding space using an image-text alignment method that consists of three different granularities: global-global alignment, global-local alignment, and local-local alignment.

In this work, we adopt a dual-stream approach to extract facial and textual features from pre-trained encoders. We use a visual semantic alignment loss, known as DAMSM [36], to align the visual and word tokens locally and globally. We also employ CMPC loss [40] to enhance the discriminative power of the features. Finally, for fine-grained cross-modal interaction, we design CFAM.

3. Framework

An overview of our proposed framework is depicted in Figure 2. The framework comprises three modules: a feature extraction module consisting of an ArcFace FR module, a contextual feature aggregation module, CFAM, and a refined pre-trained text encoder, TFRM.

3.1. Facial Feature Extraction

We first employ the ArcFace model [2] as a feature extractor to extract the facial features from the input image. Specifically, we choose ResNet18-IR [8, 2] for the backbone of the ArcFace model, which was pre-trained on the MS1MV3 dataset [3]. Here, we modify the ResNet18-IR architecture by replacing the global average-pooling layer with a fully connected layer. The output of the fully connected layer is a 512-dimensional feature vector, which is considered as the global features $\mathbf{v} \in \mathbb{R}^{512}$ of the input image, as it contains high-level visual information. We extract the local features of the image $I \in \mathbb{R}^{256\times14\times14}$ from the output of the third IR block. The size of the input image is $3 \times 112 \times 112$. We further employ CGFR on the AdaFace

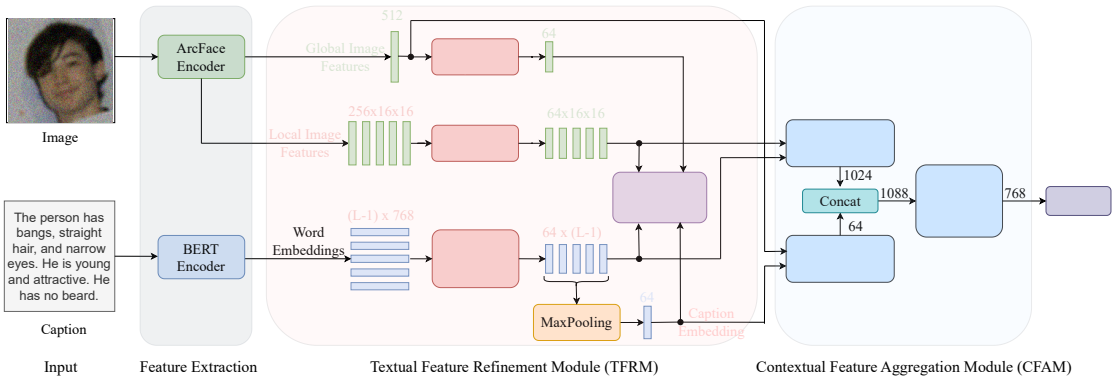

Figure 2. An overview of our proposed CGFR framework: it contains an ArcFace FR module and a pre-trained text encoder for extracting the facial features and textual embeddings from the input image-caption pair, respectively. First, TFRM updates contextual associations of the text embedding by finetuning the text encoder using the state-of-the-art DAMSM loss [36] and a cross-modal projection loss [40]. Next, CFAM fuses the facial features with textual embeddings through cross-attention at both the word and caption-level of granularities.

model [14]. Here, the backbone ResNet18-IR is similar to the ArcFace model; however, it was pre-trained on the WebFace4M dataset [42]. In contrast to ArcFace, the input is a BGR image.

3.2. Textual Feature Extraction

3.2.1 BiLSTM

Most of the works in the literature usually employ a long short-term memory network (LSTM) [10] as an encoder to extract text embeddings from natural language descriptions [35, 36]. Therefore, in this work, as a baseline, we use a bidirectional LSTM (BiLSTM) [27] as a text encoder to extract semantic vectors from the input captions. The BiLSTM encoder encodes the input caption as a matrix of $W \in \mathbb{R}^{L \times D}$. Here, D denotes the dimension of the word vector, and L denotes the maximum number of words in a caption. In our experiment, for the BiLSTM encoder, we consider a maximum of 18 words per caption, and the dimension of the word embedding is 256. Therefore, for an input caption of L words, the word embeddings are $[\mathbf{w_1}, \mathbf{w_2}, \cdots \mathbf{w_L}]$, where $\mathbf{w_L} \in \mathbb{R}^D$ is the caption embedding.

3.2.2 BERT

One of the limitations of traditional word embedding (such as word2vec) is that they only have one context-independent embedding for each word. Devlin *et al*. [4] introduced BERT, a deep bidirectional encoder that considers the context of a word for each occurrence. In this work, we adopt a pre-trained BERT-base model with 12 encoder layers, each having 12 attention heads. It obtains the contextual embedding of each word by exploiting all the words in the caption. Furthermore, in addition to the input tokens, we add a $[CLS]$ token at the beginning and a $[SEP]$ token at the end of each sentence in the caption. The maximum length of the token sequence, L, is set to 21. Additional $[PAD]$ tokens are added for short-length captions after the last $[SEP]$ token. Extra tokens are truncated if the length of the input tokens is higher than L. Therefore, the input to the BERT-base model looks like this:

$$[CLS],\ w_2,\ w_3, \cdots,\ w_{L-3},\ [SEP],\ [PAD],\ [PAD], \cdots$$

The output of the BERT layer gives a word matrix, $W \in \mathbb{R}^{L \times 768}$, where each contextualized token has an embedding of 768 dimensions. Here, the first token, $[CLS]$, is a classification token that represents the global embedding of the caption. The remaining $L-1$ tokens represent the contextualized word embeddings. In addition to BERT-base, we also experimented with other variants of BERT such as BERT-large, DistilBERT-base [26], and RoBERTa-base [20]. However, in our experiments, we found that the performance of these variants is almost the same.

3.3. Textual Feature Refinement Module

In this subsection, we briefly describe the proposed textual feature refinement module (TFRM). Because our text encoder was pre-trained with objectives that are totally different from ours and it creates embeddings that are unaligned to the image features, we need to refine the textual features. We design TFRM to realign the word and caption embedding with visual information. As shown in Figure 2, our TFRM consists of a convolution-based projection for text embeddings, a projection head for local image features, and a module to implement the visual-semantic alignment loss, DAMSM, and a cross-modal projection classification loss.

3.3.1 Projection Heads

Convolution-based Projection: As a caption has a natural order of word sequences, it is useful to extract not only

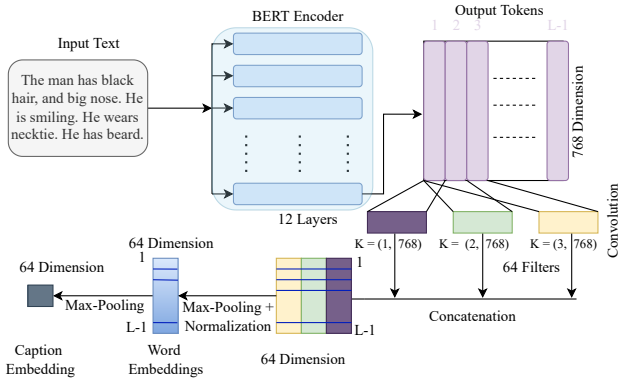

Figure 3. The proposed convolution-based projection for creating the word embeddings and global embedding for the caption. 2D-convolutions with three different kernel sizes are applied to the output representations of the BERT encoder to extract both the word- and phrase-level information.

word-level features but also phrase-level features. Thus, we apply a 2D-convolution to the output of the BERT sequence to extract both word-level and phrase-level information from the input caption. The first dimension of the kernel size K is set to 1, 2, and 3 to project uni-gram, bi-gram, and tri-gram word sequences, respectively. For $K = 2$ and 3, the word representations, W, are appropriately padded to maintain the fixed length of the sequence. All of these convolutions have a total of 64 filters with a stride of 1. Next, we apply the max-pooling operation followed by an L_2 normalization across the outputs of the convolutions to generate the word embeddings, $\mathbb{R}^{(L-1)\times 64}$. Figure 3 illustrates the proposed scheme for creating word embeddings from the output of BERT encoder.

Caption Embedding: There are multiple ways of creating the global embedding for the input caption, $\mathbf{c} \in \mathbb{R}^{64}$. One common way to create caption embedding is to employ a linear projection followed by a batch normalization [12] on the $[CLS]$ token of the BERT output layer. We can also create the caption embedding by applying the max-pooling operation across the word embeddings of the convolution-based projection followed by an L_2 normalization. However, we achieved better results from the embeddings which was created by the later scheme. Figure 3 also depicts the scheme.

Projection Head for Image Features: We need to project the local image features I, into 64 dimension which is the optimal dimension of the word embeddings. So, we design a projection head which consists of a 1×1 convolution with 64 filters and a Leaky ReLU [34] for non-linearity.

3.3.2 Objective Function

DAMSM Loss: AttnGAN [36] introduced DAMSM loss to align image-caption pair by using word-level and caption-level attention. Let (W, I) denote an image-caption pair, where $W \in \mathbb{R}^{L\times D}$ represents the word embeddings, and $I \in \mathbb{R}^{H\times W\times D}$ represents the transposed local image features. Then, we apply DAMSM loss [36] to perform cross-modal contrastive learning between image-caption pair. The loss actually minimizes the negative log posterior probability of the similarity scores between the image-caption pair.

Cross-Modal Projection Classification Loss: In order to produce discriminative textual features, we also apply a cross-modal projection classification (CMPC) [40] loss. This loss first tries to project the representations from one modality onto the corresponding features from another modality and then classify them using normalized softmax loss. The input to the CMPC is the global image features, $\mathbf{v}$, and caption embeddings, $\mathbf{c}$. First, the image features are projected onto the normalized text embeddings, $\bar{\mathbf{c}}$. Therefore, the normalized softmax loss for the image features, L_{ipt}, is given by:

$$L_{ipt} = \frac{1}{N}\sum_i -log(\frac{\exp(W_{y_i}^T \hat{\mathbf{v}}_i)}{\sum_j \exp(W_j^T \hat{\mathbf{v}}_i)}). \quad (1)$$

Here, $\hat{\mathbf{v}}_i = \mathbf{v}_i^T \bar{\mathbf{c}}_i . \bar{\mathbf{c}}_i$ denotes the vector projection of the image features. Now, let's project the text embeddings onto the normalized image features, $\bar{\mathbf{v}}$. Therefore, the text classification loss function, L_{tpi}, is given by:

$$L_{tpi} = \frac{1}{N}\sum_i -log(\frac{\exp(W_{y_i}^T \hat{\mathbf{c}}_i)}{\sum_j \exp(W_j^T \hat{\mathbf{c}}_i)}). \quad (2)$$

Here, $\hat{\mathbf{c}}_i = \mathbf{c}_i^T \bar{\mathbf{v}}_i \bar{\mathbf{v}}_i$ denotes the vector projection of the textual features. The total CMPC loss is the summation of the two losses, as defined by Eq. 1 and Eq. 2.

Full Objective: Our overall loss function is the weighted combination of the DAMSM and CMPC losses:

$$L_{loss} = \lambda_1 L_{DAMSM} + \lambda_2 L_{CMPC} \quad (3)$$

where, λ_1 and λ_2 are the hyperparameters that control the DAMSM and CMPC losses, respectively.

3.4. Contextual Feature Aggregation Module

Recent works suggest that introducing global image-caption associations in addition to fine-grained word-region interaction can lead to a more effective cross-modal fusion [24]. Therefore, it is equally important to enforce both

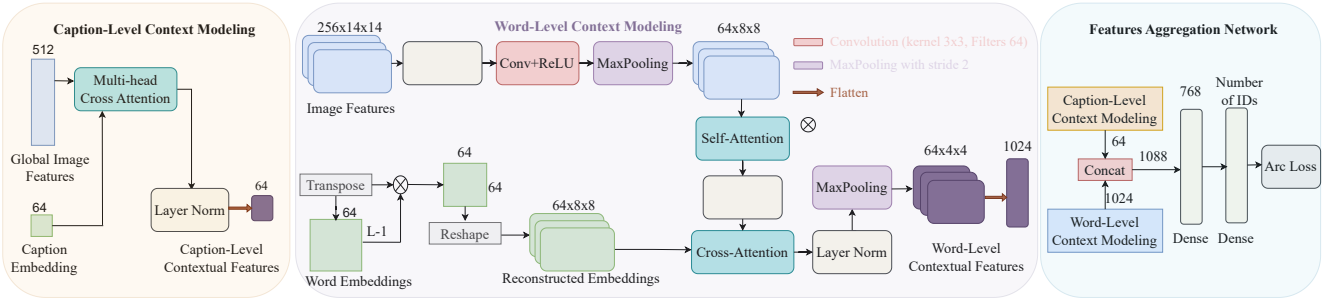

Figure 4. The block diagram of the proposed contextual feature aggregation module. It applies cross-modal feature interaction on both word and caption levels using an attention-guided mechanism. The module consists of three networks. The first network, caption-level context modeling, produces a 64-dimensional global context-enhanced features whereas the second network produces a 1024-dimensional regional context-enhanced features. The final network aggregates the contextual features and finds an optimal representation of it.

word-region interaction and global image-caption associations. In this study, we propose a contextual feature aggregation module (CFAM) that applies cross-modal feature interactions in two different granularities: word and caption. The block diagram of the proposed CFAM is illustrated in Figure 4.

3.4.1 Linear Fusion

First, we concatenate the global image features $\mathbf{v} \in \mathbb{R}^{512}$ and the caption embeddings $\mathbf{c} \in \mathbb{R}^{64}$ from the convolution-based projected head. Thus, we have a joint 576-dimensional multimodal representation. We then apply a fully connected (FC) layer. This network serves as a fusion scheme for our baseline approach.

3.4.2 Word-Level Context Modeling

In this network, we apply fine-grained cross-modal interactions between local image features and word embeddings. Here, we use word embeddings as cues to attend to the local image features extracted from the ArcFace FR module. We also experimented with using image features as cues to attend to words. However, that did not improve the performance, as words in a caption contain more abstract information than image regions. Figure 4 illustrates the word-level context modeling.

The inputs to the network are the word embeddings matrix, $W \in \mathbb{R}^{L \times 64}$, and local image features $I \in \mathbb{R}^{256 \times 14 \times 14}$. Batch normalization [12] is applied to the image features, before feeding it to a convolution layer of 64 filters with a kernel of size 3, and padding of 2. A max-pooling layer with a stride of size 2 is applied to the features map to reduce the spatial size to $64 \times 8 \times 8$. Next, a self-attention layer with a $scale = 0.5$ is applied to increase the intra-modality relationship among the local image features, followed by layer normalization [1].

Thus, due to the application of self-attention, each image region now contains information about the whole image. In the self-attention layer, the $keys$, $queries$, and $values$ are learned from 1×1 convolutions. However, the number of filters in 1×1 convolutions for projecting key and $query$ are the multiplication of a $scale$ factor of the number filters of the 1×1 convolution for learning $values$. Note that the application of normalization and self-attention in this network, as analyzed in Table 3, is very crucial.

Contrary to image features, word embeddings W, have different dimensions. Therefore, we, first, calculate the correlation of the word embeddings matrix, $W^T W \in \mathbb{R}^{64 \times 64}$. Next, we reshape the embeddings matrix to size $64 \times 8 \times 8$. Similar to image features, we also experimented to implement self-attention to the reconstructed word features, but that does not improve the performance. The reason for this could be that as textual features are extracted from transformer-based BERT architectures, the intra-modal relationship among the features is already high. Afterward, the word embeddings and image features are fed into a cross-attention scheme to increase the inter-modality relationship. Here, the $queries$ are learned from the word embeddings matrix, and $keys$ and $values$ are learned from the image features using 1×1 convolutions with a $scale$ of 0.5. Finally, after applying another max-pooling layer, we flatten the feature matrix to produce a 1024-dimensional output.

3.4.3 Caption-Level Context Modeling

Similar to the word level of granularity, we take the caption embedding as cues to attend to the global image features. Multi-head cross-attention has been employed to explore inter-modal associations between global image features and caption embedding. First, we reshape the global features into $\mathbf{v} \in \mathbb{R}^{8 \times 64}$. Then, we calculate the $queries$ from caption embedding, $\mathbf{c} \in \mathbb{R}^{64}$ and the $keys$, and $values$ from global features $\mathbf{v}$ using linear projection. The output of the cross-attention is a 64-dimensional vector, which is followed by a layer normalization [1].

Table 1. The 1:1 verification and 1:N identification (Rank-1) results of our CGFR framework with ArcFace trained on the MMCelebA dataset. The top row represents the results of ArcFace when pre-trained on MS1MV3 dataset [3].

Architectures	ROC Curve		TPR@FPR		Id(%)
	AUC	EER	1e-4	1e-3	Rank-1
Pre-trained ArcFace [2]	85.27	23.48	16.75	25.73	17.56
Baseline	93.98	13.50	21.92	31.28	38.78
CGFR	**98.51**	**6.65**	**66.83**	**68.28**	**67.65**

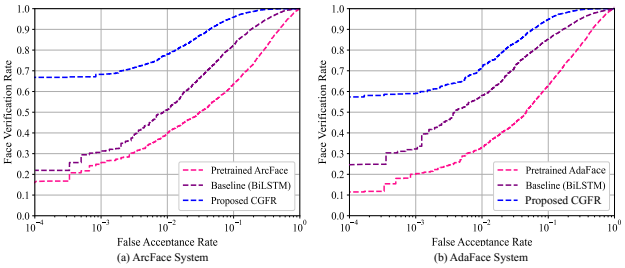

Figure 5. ROC curves of 1:1 verification protocol of the proposed CGFR framework with (a) ArcFace, (b) AdaFace FR module.

3.4.4 Features Aggregation Network

At the final stage of CFAM, we aggregate the contextualized features from the word-level CM and caption-level CM. Finally, a dense layer learns the optimal representation in a joint multimodal feature space. In our experiment, we found that the optimal dimension of the dense layer is 768. We also experimented to implement the CFAM module on the textual features extracted from the BiLSTM text encoder. However, it does not perform well as we failed to obtain the contextual embeddings from the BiLSTM encoder.

3.5. Training Strategy

We train our proposed framework in two phases. First, we train the TFRM module to update the contextual embeddings of the text encoder using the objective function mentioned in Equation 3. We finetune the BERT encoder for only 4 epochs and use a mini-batch AdamW optimizer [23] with a weight decay of 0.02. The learning rate is initialized to 0.00001 and is warmed up to 0.0001 after 2,000 training iterations. We then decrease it using the cosine decay strategy [22] to 0.00001. The batch size is set to 16. For the projection head of both visual and textual streams, we employ the Adam optimizer [15] with $\beta_1 = 0.5$ and $\beta_2 = 0.99$. The initial learning rate, in this case, is set to 0.001. In the second phase, we train the whole framework end-to-end for 24 more epochs. However, the text encoder and the projection head were trained with a similar setup to the first phase except with a lower learning rate. Note that, in all the phases, the parameters of the FR module were fixed.

Table 2. The 1:1 verification and 1:N identification (Rank-1) results of our CGFR framework with AdaFace trained on the MMCelebA dataset. The top row represent the results of AdaFace when pre-trained on the WebFace4M dataset [42].

Architectures	ROC Curve		TPR@FPR		Id(%)
	AUC	EER	1e-4	1e-3	Rank-1
Pre-trained AdaFace [14]	85.55	22.88	11.46	20.00	8.45
Baseline	93.97	12.88	24.28	33.00	22.55
CGFR	**98.10**	**7.52**	**58.08**	**59.12**	**53.23**

4. Experiments

4.1. Dataset

The Mulit-Modal CelebA-HQ [33] (MMCelebA) is a large-scale text-to-face dataset, originally built for face image generation and facial attributes editing. It has a total of 30,000 high-resolution face images from the CelebA-HQ dataset [21]. The dataset is split between 24,000 training images and 6,000 test images. Each image has 10 auto-generated captions from a total of 38 facial attributes.

4.2. Preprocessing

First, we implement standard data augmentation techniques such as random sub-sampling, color jittering, horizontal flipping, rotation, and Gaussian noise to degrade the image quality of the MMCelebA dataset. Then, we resize all the images to 3 x 112 x 112. Sample preprocessed images are shown in the Figure 1. The top row of Table 1 and Table 2 represent the performance of the pre-trained ArcFace and AdaFace models on this preprocessed dataset, respectively. From these tables, we observe that the performance of both the pre-trained ArcFace and AdaFace models substantially degraded due to the poor generalization on the low-quality images, which adversely affects their facial analysis procedure [41].

4.3. Implementation

We implemented our architecture using two NVIDIA Titan RTX GPUs. In our experiment, we empirically set the hyper-parameters in Equations 3 as follows: λ_1 = 1, and λ_2 = 0.5. Since we employ pre-trained encoders, training the proposed framework is very fast. Finetuning BERT for 4 epochs takes approximately 80 minutes on the MMCelebA dataset while training the whole network end-to-end takes 8 hours. Also, due to the parallel strategy of our proposed framework, the model has a very low time complexity during inference. The inference time, which requires only one forward process, is 220ms for an image-caption pair.

4.4. Performance Evaluation

ArcFace System: We compare our proposed CGFR to the pre-trained ArcFace and the baseline approach, as shown

Table 3. Ablation experiments of different networks on the CFAM module. Experimental results verifies the notion of fusing cross-modal features at multiple granularities improves 1:1 VR(%).

Modules	ROC Curve		TPR@FPR	
	AUC	EER	1e-4	1e-3
w/o modules	89.96	18.27	15.95	21.42
Word (w/o Norm)	86.36	22.27	8.63	20.63
Word (w/o SA)	96.30	10.42	27.02	33.13
Word (SA+Norm)	96.86	9.88	53.42	54.45
Word + Caption	97.22	9.38	63.75	64.42
Word + Caption + FAN	**98.51**	**6.65**	**66.83**	**68.28**

in Table 1. Our baseline is a dual-stream model employing a BiLSTM text encoder with a linear fusion. In the 1:1 verification protocol, the proposed CGFR achieves the highest verification rates (VR). It improved the pre-trained ArcFace by 71.68% and the baseline by 50.74% on the equal error rate (EER) metric. Also, using the true positive rate (TPR) and false positive rate (FPR) metrics, as illustrated in Figure 5(a), our proposed CGFR improves the VR(%) by a significant margin. In particular, as compared to the pre-trained ArcFace model, our framework boosts TPR(@FPR=1e-4) from 16.75% to 66.83%.

Similarly, when compared to the baseline approach, our framework improves the TPR(@FPR=1e-4) from 21.92% to 66.83% and TPR(@FPR=1e-3) from 31.28% to 68.28%. Furthermore, in the 1:N identification protocol, the proposed CFGR secures an improvement of 74.44% and 285.25% on Rank-1 identification accuracy over baseline and pre-trained ArcFace, respectively. Therefore, as the results show, the ArcFace FR module, which performs poorly due to low quality and noise, could be significantly improved using natural language supervision.

AdaFace System: In Table 2, we conduct further experiments to evaluate the performance of our CGFR framework with an AdaFace FR module. As illustrated in Figure 5(b), our framework significantly improves the VR(%) over the baseline and pre-trained AdaFace. It improves the pre-trained AdaFace by 67.13% and the baseline by 41.61% on the EER metric. Also, in 1:1 verification protocol, under the evaluation metric of TPR(@FPR=1e-4), our framework boosts the performance of pre-trained AdaFace from 11.46% to 58.08% and TPR(@FPR=1e-3) from 20.00% to 59.12%. Furthermore, as reported in Table 2, the Rank-1 identification accuracy of our CGFR framework improves by 136.05% over the baseline. Thus, the VR (%) of the above-mentioned experiments proves the effectiveness and generalizability of the proposed framework.

4.5. Analysis of CFAM

We design an ablation experiment to evaluate the effectiveness of the proposed CFAM module. Specifically,

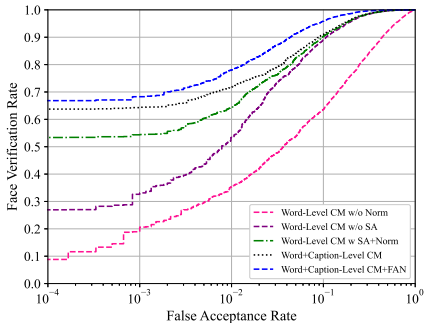

Figure 6. Face verification evaluation on different modules of the proposed CFAM using ROC curves. Verification rates (%) illustrates the need for fusing contextual features in both word-level and caption-level granularities.

we analyze the role of individual granularities and attention schemes. Table 3 demonstrates that a fusion scheme without any granularity decreases the VR(%) and proves the need for fusing contextual features at multiple granularities. In fact, under the evaluation metric of TPR, word-level contextual modeling (CM) increases the performance by 234.92% (@FPR=1e-4) over the simple concatenation of multimodal features. However, the choice of adding normalization [12, 1] and self-attention are crucial to the performance of this module. We observe a drastic performance drop of 12.15% in AUC without normalization (one batch norm [12] and two-layer norm [1]). Also, adding self-attention to the image features reduces the EER from 10.42% to 9.88%.

We also observe that the fusion of word-level and caption-level CM improves the VR(%) by 5.06% on EER and 19.34% on TPR@FPR=1e-4 compared to word-level CM. Furthermore, the ablation study shows that the implementation of the feature aggregation network further boosts the VR(%), improving TPR from 63.75% to 66.83% (@FPR=1e-4). Figure 6 depicts the performance comparison of these networks on ROC curves. Figure 6 illustrates that the proposed CFAM, with both CM networks and the feature aggregation network, achieves the highest VR(%), proving the effectiveness of applying fine-grain word-region and image-caption interaction.

5. Conclusion

We have introduced a new framework, called the caption-guided face recognition (CGFR) model, to improve the performance of FR systems using textual descriptions. Our framework is based on a dual-stream model with a textual feature refinement module (TFRM), and a contextual feature aggregation module (CFAM). CFAM applies fine-grained cross-modal feature interaction at multiple granularities using cross-attention. In contrast, TFRM helps the

framework to learn an effective joint multimodal embedding space by realigning the text embeddings with visual features. Our CGFR has significantly improved the performance of two FR models. It has also enhanced the robustness and reliability of the FR systems by offering higher resistance to spoofing attacks. In the future, we aim to employ large-scale face image-caption pair datasets to assess the generalizability of our proposed method.

References

[1] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4685–4694, 2019.

[3] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi. Lightweight face recognition challenge. In *IEEE International Conference on Computer Vision Workshops*, pages 2638–2646, 2019.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019.

[5] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez. Facial soft biometrics for recognition in the wild: Recent works, annotation, and COTS evaluation. *IEEE Transactions on Information Forensics and Security*, 13(8):2001–2014, 2018.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, and et al. Generative adversarial nets. In *Neural Information Processing Systems*, 2014.

[7] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Face recognition from caption-based supervision. *International Journal of Computer Vision*, 96(1):64–82, 2012.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[9] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. AttGAN: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019.

[10] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[11] Q. Huang, L. Yang, H. Huang, T. Wu, and D. Lin. Caption-supervised face recognition: Training a state-of-the-art face model without manual annotation. In *European Conference on Computer Vision*, pages 139–155, 2020.

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[13] D. B. J. Lu, D. Parikh, and S. Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[14] M. Kim, A. K. Jain, and X. Liu. AdaFace: Quality adaptive margin for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 18750–18759, 2022.

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He. Stacked cross attention for image-text matching. In *European conference on computer vision*, pages 201–216, 2018.

[17] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[18] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang. Identity-aware textual-visual matching with latent co-attention. In *IEEE International Conference on Computer Vision*, pages 1908–1917, 2017.

[19] W. Li, C. Gao, G. Niu, X. Xiao, H. Liu, J. Liu, H. Wu, and H. Wang. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 2592–2607, 2021.

[20] Y. Liu, M. Ott, N. G., J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyano. RoBERTa: A robustly optimized bert pretraining approach. *CoRR, abs/1907.11692*, 2019.

[21] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[22] I. Loshchilov and F. Hutter. SGDR: stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

[23] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[24] K. Niu, Y. Huang, W. Ouyang, and L. Wang. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing*, 29:5542–5556, 2020.

[25] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, and et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[26] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[27] M. Schuster and K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[28] H. Shi, H. Li, F. Meng, and Q. Wu. Key-word-aware network for referring expression image segmentation. In *Proceedings of the European Conference on Computer Vision*, 2018.

[29] J. Sun, Q. Deng, Q. Li, M. Sun, M. Ren, and Z. Sun. AnyFace: Free-style text-to-face synthesis and manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 18687–18696, 2022.

[30] P. Tome, J. Fierrez, R. Vera-Rodriguez, and M. S. Nixon. Soft biometrics and their application in person recognition at a distance. *IEEE Transactions on Information Forensics and Security*, 9(3):464–475, 2014.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and et al. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[32] T. Wang, T. Zhang, and B. Lovell. Faces à la Carte: Text-to-face generation via attribute disentanglement. In *2021 IEEE Winter Conference on Applications of Computer Vision*, pages 3379–3387, 2021.

[33] W. Xia, Y. Yang, J.-H. Xue, and B. Wu. TediGAN: Text-guided diverse face image generation and manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2256–2265, 2021.

[34] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[35] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, and et al. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[36] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018.

[37] L. Ye, M. Rochan, Z. Liu, and Y. Wang. Cross-modal self-attention network for referring image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10494–10503, 2019.

[38] X. Yin, Y. Tai, Y. Huang, and X. Liu. FAN: Feature adaptation network for surveillance face recognition and normalization. In *Asian Conference on Computer Vision*, 2020.

[39] H. Zhang, J. R. Beveridge, B. A. Draper, and P. J. Phillips. On the effectiveness of soft biometrics for increasing face verification rates. *Computer Vision and Image Understanding*, 137:50–62, 2015.

[40] Y. Zhang and H. Lu. Deep cross-modal projection learning for image-text matching. In *European conference on computer vision (ECCV)*, pages 686–701, 2018.

[41] S. Zhu, S. Liu, C. C. Loy, and X. Tang. Deep cascaded bi-network for face hallucination. In *European conference on computer vision*, pages 614–630, 2016.

[42] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, and et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *IEE Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021.