

Transformer With Decoupled Self-Attention Regularization for Age-Unbiased Facial Expression Recognition

Jaeil Park , Sung-Bae Cho  *

Dept. of Computer Science, Yonsei University, Seoul 03722 (South Korea)

* Corresponding author: wodlf603@yonsei.ac.kr (J. Park), sbcho@yonsei.ac.kr (S. B. Cho).

Received 15 April 2025 | Accepted 20 July 2026 | Published 28 August 2026



ABSTRACT

Facial expression recognition (FER) is a challenging task that involves inferring human emotions from facial images entangled with various attributes. Although FER models based on Transformer have recently reported impressive performance, two major types of bias continue to degrade accuracy—namely, age-related facial attributes such as wrinkles and skin texture, which introduce confusion in recognizing elderly emotions, and imbalanced training data, which limits the model's ability to generalize across age groups. This paper proposes a novel bias-mitigation method that decouples and regularizes age- and emotion-related components within the self-attention mechanism of Transformer. The proposed method separately regularizes the value vectors that encode texture information correlated with age, and the query-key matrices that focus on facial landmarks crucial for emotion recognition. It encourages intra-class compactness of facial landmarks for emotion representation while minimizing age interference in features unrelated to emotion. To suppress age information and preserve emotional features, an age discriminator is employed to guide the value vectors in eliminating age cues, while an emotion classifier restores discriminative information for classification. In addition, the proposed method leverages triplet learning on the query-key space to enhance age-invariant emotion separation. Experiments conducted on four widely used FER benchmarks demonstrate that our method notably reduces age-related bias while maintaining or exceeding state-of-the-art performance.

KEYWORDS

Algorithmic Bias, Adversarial Regularization, Data Imbalance, Facial Expression Recognition, Triplet Learning.

DOI: 10.9781/ijimai.2026.2235
eLocator: e2235

I. INTRODUCTION

FACIAL expression recognition (FER) has attracted significant attention in the field of affective computing, with a wide range of applications including human-computer interaction, mental health monitoring, and social robotics [1]–[3]. FER aims to infer emotional states from facial images, a task that is inherently complex due to the presence of confounding attributes such as age, gender, and identity [5], [6]. Among these, age is particularly challenging, as it introduces visual variations (e.g., wrinkles, skin sagging, and texture degradation) that are not directly related to emotional expression but significantly influence facial appearance [8], [9].

Recent advances in deep learning, especially Transformer-based architectures, have led to remarkable improvements in FER performance [4], [10]. However, despite their capacity to model long-range dependencies and capture intricate facial patterns, these models remain susceptible to age-related bias [12]. As shown in Fig. 1, this bias manifests in two primary forms: (1) entanglement between age-

specific facial features and emotional cues, which confuses the model when interpreting expressions in elderly faces [9], and (2) imbalanced data distribution, where older age groups are underrepresented in many FER datasets [13], [14], leading to poor generalization and reduced accuracy for those populations.

Existing bias-mitigation approaches often rely on dataset balancing or adversarial learning to remove protected attributes [15]–[17]. However, these methods either require additional data or risk erasing useful discriminative features. On the other hand, we observe that different components of the Transformer attention mechanism implicitly encode different types of information: the query-key (Q-K) matrices are primarily related to spatial structure and facial landmarks, which are crucial to emotion representation, while the value (V) vectors capture textural details, which are more susceptible to age-related variations. This structural property motivates us to address age bias by decoupling and regularizing these components separately.

Please cite this article as:

J. Park, S. B. Cho, Transformer With Decoupled Self-Attention Regularization for Age-Unbiased Facial Expression Recognition, *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 10, no. 1, e2235, 2026, <http://doi.org/10.9781/ijimai.2026.2235>

II. RELATED WORKS

A. FER Models and Their Bias

Recent advancements in FER have been primarily driven by sophisticated machine learning and computer vision techniques. These developments have been largely based on various benchmark datasets, such as the Real-world Affective Faces Database (RAF-DB) [18], AffectNet [19], and FER2013 [20], which serve as fundamental resources for training and testing FER models.

Traditional FER models are trained on these datasets. The upper section of Table I provides a summary of existing FER models. With the advancement of deep learning, models based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been extensively studied, leading to the development of FER models that leverage these architectures. These methods have successfully captured complex emotional expressions, significantly improving FER accuracy. For example, Shi et al. introduced a feature arrangement and a de-albino module to recover lost feature information [21]. Pham et al. proposed an attention residual network incorporating landmark masks for facial emotion recognition [22]. These models demonstrate the importance of preserving and utilizing key information for FER, such as facial landmarks.

Furthermore, with the rise of transformer-based architectures, vision transformers (ViTs) have been explored in the image domain. Given their high accuracy across various domains, numerous studies have adapted ViTs for FER, achieving *state-of-the-art* performance. For instance, Xue et al. implemented a ViT with attention pooling to effectively reduce the model's focus on irrelevant noisy regions [4]. Mao et al. designed a two-stream model based on image landmarks with multi-scale feature extraction and cross-fusion, achieving high accuracy [10]. Chen et al. proposed a model that enhances dynamic image recognition by utilizing facial landmark features extracted from static images [23]. Wang et al. introduced a method leveraging intra-class and inter-class cross-similarity to extract intrinsic information, further improving ViT performance [24].

However, bias and fairness in FER have emerged as critical concerns. Studies by Gou et al. [25] and Howard et al. [26] highlighted biases in age, gender, and ethnicity recognition, emphasizing the need for unbiased algorithms and diverse training datasets. This aspect of research underlines the importance of ethical considerations in algorithm development.

B. Bias Mitigation for FER

In the field of FER, various methods have been explored for bias mitigation, focusing on data or model modification as shown in Table I. Gou et al. addressed aging effects in facial emotion recognition models, applying image modifications to reduce features related to age and testing different classifiers [25]. Howard et al. tackled biases in adolescent facial emotion classification, employing support vector machines for specific expressions [27]. To avoid manual labeling, Zhang et al. used a conditional generative adversarial network (cGAN) for sample generation, aiming to reduce bias through varied facial emotions [28]. Xu et al. examined both attribute-aware and attribute-unaware bias mitigation techniques, incorporating data augmentations for improved fairness [14]. Lastly, Cai et al. developed a conditional GAN architecture to remove identity-related features from facial emotions, integrating a U-Net-based generator and self-attention modules for an average emotional image, despite the cost of increased complexity and training duration [29]. Each of these methods addresses bias from different angles, highlighting the multifaceted approach required for effective bias mitigation in FER systems.

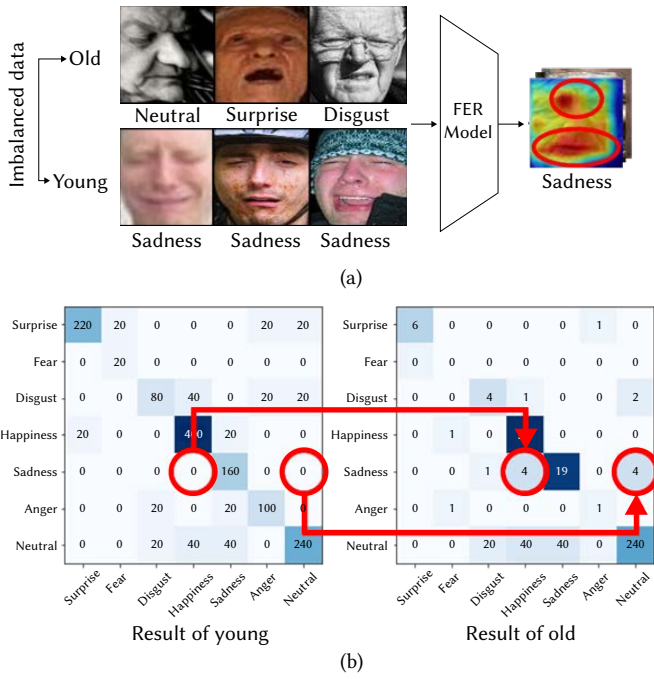


Fig. 1. (a) Age bias in FER models hinders accurate emotion recognition in older adults, with attention maps highlighting contributing features like wrinkles. (b) Confusion matrices reveal a noticeable bias in predictions for the elderly group.

In this paper, we propose a novel FER method designed to mitigate age bias while preserving emotion-related representations. This method introduces two main modules: (1) value regularization that incorporates an age discriminator to suppress age-specific information and an emotion classifier to recover useful features for expression recognition, and (2) query-key separation trained via triplet learning to enhance intra-class emotion similarity and inter-class separability, independent of age.

We conduct extensive experiments on four widely-used FER benchmarks, demonstrating that our method not only achieves competitive or superior recognition accuracy but also substantially reduces age-induced performance disparity. The main contributions of this paper can be summarized as follows:

- We identify and explicitly address the entanglement of age and emotion in Transformer-based FER models by analyzing the roles of Q-K and V components in attention.
- We propose a novel dual-regularization method that simultaneously suppresses age bias in value vectors and enhances emotion clustering in the query-key space.
- We validate our method through comprehensive experiments across multiple datasets, showing improvements in both accuracy and fairness.

The structure of this paper is as follows. Section II provides a comprehensive review of related works and discusses the limitations of existing bias mitigation methods. Section III introduces the proposed method, which performs targeted regularization on the query-key matrices and value vectors of the Transformer model. Section IV presents the experimental setup, including dataset descriptions and evaluation metrics, followed by a detailed analysis of the results. Finally, Section V summarizes the key findings and outlines directions for future research.

TABLE I. RELATED WORKS FOR FER AND ITS BIAS MITIGATION METHODS

Type	Method	Description
FER	CNN and RNN	Feature arrangement and de-albino module for ResNet-18 [21]
		Attention residual network for landmark masks [22]
	Transformer	Vision Transformer (ViT) with attention pooling [4]
		Separation of handling landmark detection and image processing [10]
Bias Mitigation for FER	Data Alteration	Aging details removal using generated mask of face images [25]
	Data Constraint	Age group constrained FER for specific age group [25]
		Age group constrained FER for detailed facial properties [27]
	Adversarial Learning	Adversarial learning with simple cGAN [28]
		Adversarial learning with concatenated features [14]
		Adversarial learning with disentangled feature [14]
		Adversarial learning with identity removal cGAN [29]
	Few-shot	Reweighting-based methods to improve the generalization [30]
		Applying 10 meta-learning and few-shot learning methods [31]

Moreover, since human emotions tend to be expressed in a biased manner due to social and cultural influences, FER datasets inherently exhibit data imbalance. Training on imbalanced data leads to model bias. Bauder and Khoshgoftaar pointed out that when class imbalance is extreme, a classifier is likely to predict most instances as belonging to the majority class, leading to a seemingly high overall accuracy [32]. However, when the prediction performance of minority classes is also crucial, such an approach becomes inefficient. Thabtah et al. argued that addressing data imbalance can improve accuracy [33]. Hoitsma et al. classified bias into implicit and explicit categories and emphasized the importance of addressing them to mitigate bias effectively [34].

Various studies have explored solutions to data imbalance bias across different domains. Ren et al. proposed balanced softmax to mitigate class imbalance caused by long-tailed data distribution, achieving improved accuracy compared to conventional methods [35]. Wei et al. employed probabilistic selection of minority class samples and iterative self-training to improve accuracy [36]. Few-shot learning approaches have also been investigated to address data imbalance bias. Ren et al. leveraged reweighting-based methods to improve the generalization capability of models [35]. Ochal et al. demonstrated that few-shot learning techniques can be highly effective, achieving state-of-the-art performance [31]. These studies highlight the significant impact of data imbalance bias on accuracy and the importance of addressing it. Furthermore, they demonstrate that few-shot learning methods can be highly effective in mitigating bias.

In the FER domain, efforts have been made to address data imbalance as well. Wasi et al. proposed a facial geometry-aware anchor-based reliability balancing module for ViT to mitigate the impact of imbalanced data distribution, achieving superior accuracy compared to other models [37]. Sreenivas et al. introduced the DIFC module to improve performance by addressing data imbalance in the unsupervised domain adaptation setting [38].

However, these bias mitigation techniques often require complex preprocessing steps that are difficult to apply to general-purpose models. Moreover, the generation process or adversarial learning can lead to direct information loss, paradoxically degrading model accuracy. This presents a major challenge in achieving effective bias removal while preserving essential information for accurate recognition. In this paper, we propose a method that regularizes age-related and data imbalance-related attributes within the self-attention mechanism of Transformer models, specifically targeting their respective representations. This method effectively mitigates age-induced bias and improves model fairness. Experimental results demonstrate notable performance gains, validating the effectiveness of the proposed method.

III. METHODOLOGY

A. Bias of FER and Self-attention of FER Models

This section provides a brief overview of the self-attention mechanism in Transformer [39], and its relationship with an age bias. The transformer employs an encoder-decoder architecture for sequence-to-sequence learning. In the context of images, for example, the vision transformer (ViT) [40] applies this approach by segmenting images into patches and transforming these sequences into other sequences, demonstrating strong performance with self-supervised learning. During this process, the Transformer's encoder transforms the input image patches into internal hidden state vectors, while the decoder converts these vectors into an output sequence.

Each of the encoder and decoder is composed of multiple layers of computational functions. Within each layer of the encoder, there are two sequential sub-layers: multi-head self-attention and a position-wise feed-forward network. Beyond the multi-layer architecture of the encoder, the decoder includes an additional attention layer, so-called encoder-decoder attention, which aids the model in focusing on the internal states of the encoder. Both the encoder and decoder layers also incorporate skip-connection and layer normalization in their computational pipeline. The overall structure of the encoder and decoder consists of stacks of L identical layers, including an embedding layer.

The computational process can be summarized with mathematical notations. The size of the hidden state vector is denoted as d_{model} and the number of heads in multi-head self-attention is represented as h . The dimensions of the query, key, and value vectors are $d_q = d_k = d_v = d_{model}/h$.

At each attention head, encoded word vectors are transformed into three matrices: a query matrix $Q \in R^{t \times d_q}$, a key matrix $K \in R^{t \times d_k}$, and a value matrix $V \in R^{t \times d_v}$. This transformation occurs h times, resulting in $h \times 3$ matrices. These are used to calculate the attention-weighted combination of value vectors as follows:

$$A(Q, K) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_{model}}}\right) \quad (1)$$

$$\text{Attention}(Q, K, V) = A(Q, K)V \quad (2)$$

$$\text{MultiHeadAttention} = \text{Concat}(H_1, \dots, H_h)W^O \quad (3)$$

where $H_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, and W_i^Q , W_i^K , and W_i^V are the linear transformation matrices at the i -th head. A is called an attention map which represents relations of each vector as a weight of similarity [39], [41].

Fig. 2 (a) illustrates the attention map in FER. The attention map, representing relations of image patches, exhibits characteristics associated with landmarks [42]. In contrast, previous research

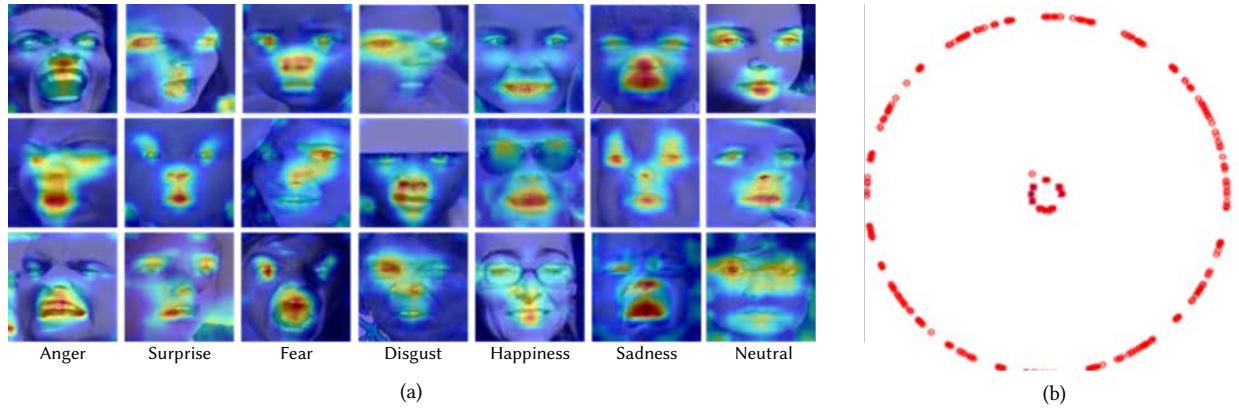


Fig. 2. Association between attention and attributes in the FER model. (a) Photo extracted with the attention map of FER. It can be observed that there is a deep association with landmarks. (b) tSNE results of the value vector and the age attribute.

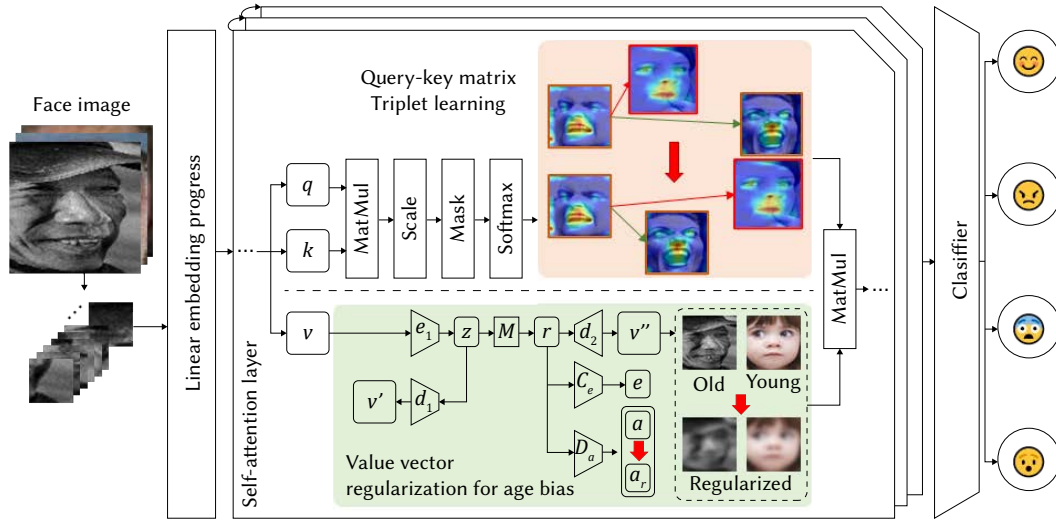


Fig. 3. The overall architecture of the proposed model.

has shown that value vectors represent local information [43]. Additionally, studies have found a high correlation between facial age information and local information in images [44], [45]. Fig. 2 (b) visualizes the relationship between the distribution of the value vector and age, suggesting a correlation between the actual age and the value vector. Consequently, this paper aims to mitigate age bias in the value vector by exploring the relationship between the value vector and facial age information.

B. Age Bias Mitigation by Value Regularization

Fig. 3 shows the architecture of a facial emotion recognition model that mitigates bias using the value vector of self-attention. The model processes image patches as inputs, executing a Transformer. Within this architecture, the value vector of the self-attention is regularized through an autoencoder. Denoting the function that carries out this process as f_{AE} , the self-attention with the proposed autoencoder is represented as shown in equation (4).

$$Attention(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_{model}}}\right) f_{AE}(V) \quad (4)$$

The autoencoder f_{AE} aims to generate a fair representation by focusing on the latent vector regularization process in adversarial autoencoders. This process involves compressing the initial input to produce latent vectors that retain maximal information with an encoder and decoder. The distribution between the original input and the generated output is then regularized using a discriminator,

aligning their distributions. The adversarial autoencoder leverages the convergence of these distributions to effectively train the discriminator, which in turn learns the distribution of the original input.

Drawing on the idea of using latent vectors to converge and regularize data distributions, the proposed method seeks to adjust the data distributions of value vectors with age features, adopting a disentangled representation approach [46]-[48]. An age discriminator is employed to check whether the input vector originates from an older individual's image, thereby facilitating the learning of the data distribution. We achieve convergence of data distributions through a regularization process that employs the discriminator to soften the information related to age, while the decoder works to preserve the original information.

Suppose that $V = \{v_1, \dots, v_n\}$. Using an encoder e_1 and a decoder d_1 , the proposed method creates a latent vector z that contains all the information in embedded tokens v and can be decoded into v' that contains the same information as the original data. To separate the learning direction of age information removal and original information retrieval, we generate a mutated latent vector r that passes through the layer m and inherits the information from the vector. Ensuring that the mutated latent vector retains the information about the original vector as much as possible, the vectors are also decoded with a decoder d_2 and compared with v'' . With this structure, function f_{AE} can be written as equation (5):

$$f_{AE}(V) = V'' = (d_2 \circ m \circ e_1)(V) \quad (5)$$

Mutated latent vectors containing an age feature are classified by a discriminator D_a , so that the discriminator learns the distribution of two age groups of latent vectors. Subsequently, the discriminator D_a is fixed, and the age label of the mutated latent vector a is randomized as a_r . As z and r are regularized while D_a maintains to discriminate the age label of them appropriately, z and r are learned not to be distinguished by D_a . By iteratively training the discriminator and randomizing the label, z and r are learned to be disentangled for an age feature [49].

However, in cases where only a discriminator exists, removing biases can result in the loss of key information related to emotion recognition. To address this issue, we employ an emotion classifier, C_e , to prevent excessive information loss. C_e learns the emotional attributes of the latent vector before the age is randomized. After age randomization, it remains fixed along with the emotional label, thereby maintaining the emotional attributes.

For learning debiased representation, the objective function can be shown as equation (6) with information theory, where I represents the mutual information, V is the original value vector token, R is the mutated latent vector, $\mathcal{A}$ is the age attribute, E is the emotion attribute, and β is a coefficient to balance the two terms.

$$\max L = I(R; V) - \beta I(R; \mathcal{A}) \quad (6)$$

Here, maximizing $I(R; V)$ which represents the value of association between the original value vector V and the latent vector R , we can maximize the lower bound of equation (6). The lower bound of $I(R; V)$ is calculated as shown in equations (7) and (8) [50]. D_2 is the function of decoder 2, q is the encoder function, C is the classification function, p is the real distribution, m is the projection function. Let the first lower bound of $I(R; V)$ be L_{r_1} and the second be L_{r_2} .

$$I(R; V) \geq E_{v'' \sim p(v''), r \sim m(r|z), z \sim q(z|x)} [\log D_2(v''|r)] \quad (7)$$

$$I(R; V) \geq E_{r \sim m(r|z), z \sim q(z|x)} [\log C(e|r)] \quad (8)$$

At the same time, minimizing $I(R; \mathcal{A})$ which represents the value of association between the latent vector R and the protected feature attribute $\mathcal{A}$, we can also maximize the lower bound of equation (6). As an additional function m is used, we can minimize $I(R; \mathcal{A})$ by reducing the upper bound as shown in equation (9), where $s(r)$ is a variational approximation. Let this upper bound of $I(R; \mathcal{A})$ be C_1 .

$$I(R; \mathcal{A}) \leq \int m(r|z)p(z) \log \frac{m(r|z)}{s(r)} \quad (9)$$

However, it is not ideal because considering only with function $m()$ is not stable. For the tighter upper bound when assuming optimal dual regularization, $I(R; \mathcal{A})$ can be approximated as shown in equation (10), where l is positive such that $D_{KL}(p(a|r) \| h(a|r)) \leq l$ [51], where D_{KL} is KL divergence.

$$I(R; \mathcal{A}) \leq E_{p(R, \mathcal{A})} [\log p(a|r) - \log p(a)] + l \quad (10)$$

Let this upper bound of $I(R; \mathcal{A})$ be C_2 . The final objective function for bias mitigation can be formulated as equation (11).

$$\max L_{r_1} + \max L_{r_2} - \beta_1 C_1 - \beta_2 C_2 \quad (11)$$

The latent vector z and the mutated latent vector r are learned by L_{r_1} and C_1 in different directions: L_{r_1} learns z , r to store all the features of v , mainly emotion feature e , but C_1 forces v to be representation without the information of age feature. Algorithm 1 shows its training mechanism.

As shown in this algorithm, each mini-batch first samples an input x , its corresponding age label α , and a randomly mutated age attribute $\tilde{\alpha}$. The value vector v is then extracted from the attention layers and passed to the encoder. The encoder and the reconstruction layer are updated using L_{r_1} so that the latent vectors z and r retain sufficient

information from v for emotion representation learning. Next, the encoder is updated with the debiasing losses C_1 and C_2 using the mutated age attribute $\tilde{\alpha}$, which discourages the latent representation from encoding age-discriminative information. Finally, the age-related discriminator is updated using the true age label α , thereby providing an adversarial signal for the subsequent encoder update. Through this adversarial regularization, the encoder learns a latent representation that preserves emotion-related information while reducing age-related bias.

Algorithm 1. Learning a debiased latent vector

Input: Data X and corresponding age feature A

Output: Fully trained encoder q

- 1: Initialize q, D_1, D_2 ;
 - 2: **for** epochs **do**
 - 3: **for** batches **do**
 - 4: Sample $x, a, \tilde{a}$ from $X, A, U(0, 1)$ respectively;
 - 5: $v \leftarrow \text{valueVector}(x)$
 - 6: $\theta_q \leftarrow \theta_q - \eta \frac{\partial \mathcal{L}_r}{\partial \theta_q}(v, \alpha)$
 - 7: $\theta_{D_1} \leftarrow \theta_{D_1} - \eta \frac{\partial \mathcal{L}_r}{\partial \theta_{D_1}}(v, \alpha)$;
 - 8: $\theta_q \leftarrow \theta_q - \eta \frac{\partial (C_1 + C_2)}{\partial \theta_q}(v, \tilde{\alpha})$;
 - 9: $\theta_t \leftarrow \theta_t - \eta \frac{\partial (C_1 + C_2)}{\partial \theta_t}(v, \tilde{\alpha})$;
 - 10: $\theta_{D_2} \leftarrow \theta_{D_2} - \eta \frac{\partial C_2}{\partial \theta_{D_2}}(v, \alpha)$;
 - 11: **end for**
 - 12: **end for**
-

C. Triplet Learning for Query-key Matrix

To effectively address the data imbalance problem inherent in FER tasks, our method applies contrastive learning specifically to the query-key (Q-K) matrix within the Transformer architecture. The Q-K matrix is particularly suited for this purpose as it captures the spatial dependencies and structural features of facial landmarks—elements that are crucial for emotion recognition and relatively invariant to age or other demographic attributes.

In this paper, we construct training triplets by randomly selecting an anchor sample from the dataset. A positive sample is then chosen from the same class (i.e., expressing the same emotion as the anchor), while a negative sample is selected from a different class. These samples are passed through the model to obtain their corresponding attention matrices, denoted as a^a (anchor), a^p (positive), and a^n (negative). These attention matrices encapsulate the internal representation of facial landmark relationships learned by the self-attention mechanism.

To enforce class-wise separation in the learned latent space, we apply triplet loss based on the Euclidean distance between pairs of attention matrices. Specifically, the distance between the anchor and positive samples is denoted as $d_p = \|a^a - a^p\|_2$, and the distance between the anchor and negative samples as $d_n = \|a^a - a^n\|_2$. Here, regular Euclidean distance is used because it is proved to operate more stably compared to squared Euclidean distance [52]. By minimizing d_p and simultaneously maximizing d_n , the model is encouraged to group semantically similar samples closer together while pushing apart dissimilar ones, resulting in more distinct and compact class clusters in the feature space. This mechanism is formally captured by the triplet loss function, defined as equation (12).

$$\mathcal{L}_{\text{triplet}} = \mathcal{L}(a^a, a^p, a^n) = \max(d_p - d_n + \alpha_t, 0) \quad (12)$$

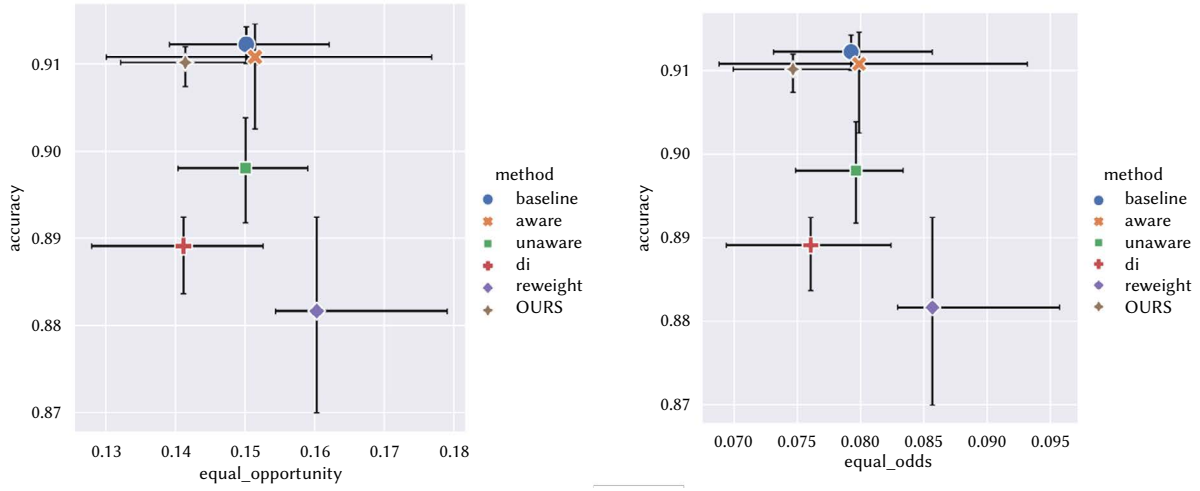


Fig. 4. Comparison of accuracy equal opportunity, and equal odds between different bias mitigation methods.

where α_i is a pre-defined margin parameter that enforces a minimum separation between positive and negative distances. Once the margin constraint $d_n \geq d_p + \alpha_i$ is satisfied, the triplet loss becomes zero and no further penalty is imposed. Thus, d_n is increased only until the margin condition is met, rather than being maximized without bound. Pushing negative samples farther than the required margin is unnecessary and may distort the latent embedding space [52], [53]. This constraint prevents the model from collapsing into a trivial solution and ensures robust class separation.

By optimizing this loss, the proposed model learns a Q-K latent space that inherently reflects emotion-specific clustering, even in the presence of data imbalance or age-related variation, thereby improving both the discriminative power and fairness of the FER model.

IV. EXPERIMENTS

A. Experimental Setting

In this paper, we employ the real-world affective faces database (RAF-DB) [18], FACES [54], AffectNet [19] and FER2013 [20] for both training and evaluation. These datasets are utilized specifically for their basic emotion subset, classifying images into seven expressions (surprise, fear, disgust, happiness, anger, sadness, and neutral) and bifurcating age groups into young (up to 69 years for RAF-DB, young/middle for FACES) and old (70 years and older for RAF-DB, old for FACES) [26]. For RAF-DB and FACES, we use the age annotations provided in the original datasets, whereas for AffectNet and FER2013 we assign age-group labels using an age classifier trained on RAF-DB. Table II summarizes the number of samples per age group and emotion category for each dataset. Age bias starts to appear from the 40s from a cognitive perspective [54], but there is a change in trend around the age of 70, leading to a significant accuracy drop [55]. Therefore, we target primarily those over 70 years old, as they are highly relevant to the age group where the problem is most pronounced. Notably, while these datasets provide labels for sensitive attributes, they are not curated with a focus on group balance. This results in an imbalanced distribution of samples across expressions and age groups in both training and testing splits, contributing to age biases in the models as indicated by Xu et al. [14].

For pre-processing, each image is resized to a resolution of 224x224 pixels and undergoes color normalization. To enhance input diversity during training, we apply trivial augment [56], implement random cropping [57], and occasionally flip images horizontally. These transformations do not alter the empirical age-label distribution of the dataset.

TABLE II. NUMBER OF SAMPLES PER AGE GROUP AND EMOTION CATEGORY FOR EACH DATASET, WHERE S1, F, D, H, S2, A, AND N DENOTE SURPRISE, FEAR, DISGUST, HAPPINESS, SADNESS, ANGER, AND NEUTRAL, RESPECTIVELY

	RAF-DB		FACES		AffectNet		FER2013	
	Young	Old	Young	Old	Young	Old	Young	Old
S1	1277	13	-	-	16120	168	3135	36
F	275	6	114	57	8015	176	4003	94
D	701	16	114	57	5140	124	428	8
H	4613	159	114	57	141326	4872	6980	235
S2	2014	68	114	57	28423	1064	4678	152
A	698	7	114	57	27810	320	3954	41
N	2427	97	114	57	77178	3098	4773	192

For training and evaluating the proposed method, the baseline model is the SOTA model POSTERv2 [10] and high-performance model with attention pooling APViT [4]. For comparison, ResNet-ARM [21], high performance FER model based on CNN, and bias mitigation methods Aware and Unaware [14] are used. The ‘Aware’ and ‘Unaware’ models follow the training protocols as outlined by Wang et al. [7]. Each model is trained and evaluated five times.

The models are developed using the Pytorch framework and optimized via the adam optimizer (with learning rate of 0.0003 and weight decay of 0.0001) coupled with a single cosine annealing schedule. The training duration spans 800 epochs. In the method incorporating an adversarial module, these additional components are optimized separately but with identical hyperparameters.

Fairness evaluation is conducted using the equal opportunity and equal odds metrics as proposed by Hardt et al. [58]. Equal opportunity is assessed through the absolute difference in the true positive rate between the two groups as in equation (13). For equal odds, we additionally compute the absolute difference in the false positive rate between the groups as in equation (14).

$$EOP = \sum_{a \in A} |TPR_{a=1} - TPR_{a=0}| \quad (13)$$

$$EOD = \sum_{a \in A} \frac{|TPR_{a=1} - TPR_{a=0}| + |FPR_{a=1} - FPR_{a=0}|}{2} \quad (14)$$

B. Bias Mitigation on FER

Overall performance is compared as shown in Table III, while Table IV presents qualitative examples of per-method predictions for older faces with a neutral ground-truth label that were misclassified as “sadness” by the baseline model. Fig. 4 shows the comparison of accuracy and bias metrics (equal opportunity and equal odds) between different bias mitigation methods. The proposed method shows the most substantial improvement in fairness metrics compared to other methods, and in

TABLE III. RESULTS WITH FOUR DATASETS. EOP IS EQUAL OPPORTUNITY AND EOD IS EQUAL ODDS

Model	RAF-DB			FACES			AffectNet			FER2013		
	Acc.	EOP	EOD	Acc.	EOP	EOD	Acc.	EOP	EOD	Acc.	EOP	EOD
APViT [4]	0.920	0.261	0.178	0.800	0.361	0.194	0.662	0.372	0.210	0.669	0.360	0.205
APViT [4]+Aware [7]	0.913	0.251	0.153	0.798	0.350	0.183	0.653	0.350	0.180	0.660	0.330	0.170
APViT [4]+Unaware [7]	0.876	0.243	0.150	0.776	0.342	0.181	0.640	0.338	0.195	0.654	0.325	0.165
POSTERv2 [10], [11]	0.922	0.323	0.169	0.819	0.347	0.184	0.678	0.205	0.675	0.675	0.355	0.200
POSTERv2 [10]+Aware [7]	0.917	0.312	0.156	0.812	0.325	0.165	0.670	0.310	0.160	0.672	0.300	0.155
POSTERv2 [10]+Unaware [7]	0.896	0.319	0.149	0.797	0.292	0.164	0.655	0.300	0.157	0.660	0.295	0.150
APViT [4]+Ours	0.922	0.257	0.136	0.812	0.318	0.158	0.685	0.290	0.150	0.678	0.275	0.145
POSTERv2 [10]+Ours	0.931	0.261	0.137	0.828	0.275	0.148	0.698	0.265	0.142	0.690	0.260	0.140

the case of ours, accuracy is also largely enhanced. When the proposed the RAF-DB dataset. While some methods demonstrate superior

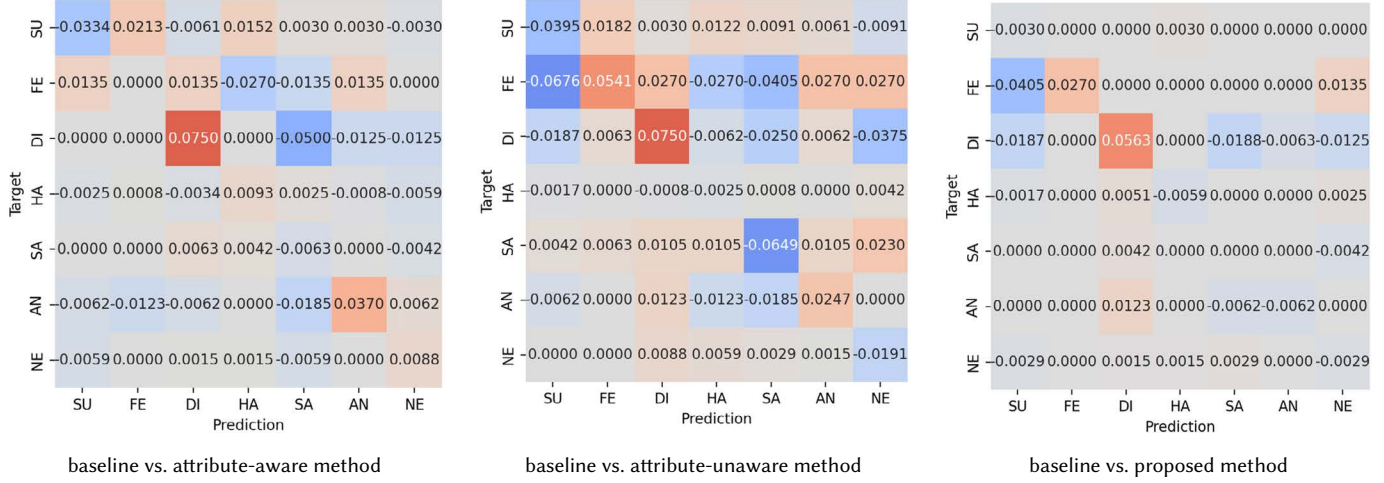


Fig. 5. Confusion matrix differences of baseline vs. aware, unaware, and proposed method.

TABLE IV. PER-METHOD PREDICTION ON OLD FACES MISCLASSIFIED AS 'SADNESS'

Target label				
	Neutral	Neutral	Neutral	Neutral
Baseline	Sadness	Sadness	Sadness	Sadness
Aware	Disgust	Sadness	Sadness	Neutral
Unaware	Neutral	Neutral	Sadness	Disgust
Ours	Neutral	Neutral	Neutral	Sadness

method is applied to APViT, there is no performance improvement in RAF-DB due to the dilution of information by attentive pooling, but a reduction in bias is observed. The methods of eliminating bias through Aware and Unaware approaches are found to decrease accuracy due to a lack of information restoration capability.

Fig. 5 highlights the differences in confusion matrices among each mitigation method relative to the baseline. Notably, the proposed method exhibits lower variance, indicating that modifications are made only to predictions on samples necessary for enhancing fairness. In contrast, as shown in Fig. 5, both the Aware and Unaware methods exhibit changes in the prediction distributions across all classes, particularly under the adversarial approach. This suggests that the new feature representations may not align with the original latent space, leading to a diminished quality of representation.

C. Age-wise Comparison

When analyzing fairness metrics across different age categories, it becomes evident that the proposed method can mitigate biases on a per-group basis. Table V and Fig. 6 elucidate the results within

performance in younger groups, ours remains comparatively effective.

TABLE V. RESULTS ON AGE-WISE COMPARISON

Age category	Model	Equal opportunity	Equal odd
Young	Baseline	0.1076	0.0574
	Aware	0.1009	0.0538
	Unaware	0.1072	0.0572
	Ours	0.1063	0.0529
Old	Baseline	0.2405	0.1278
	Aware	0.2270	0.1209
	Unaware	0.2357	0.1251
	Ours	0.1736	0.1057

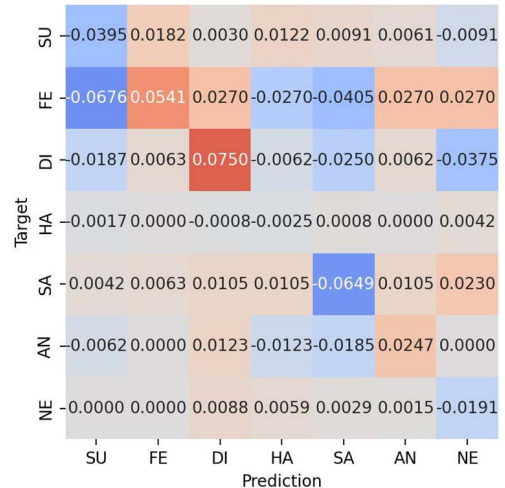


Fig. 6. Confusion matrix differences in age of baseline vs. proposed method.

The elderly category in the dataset, having the smallest sample size, is more prone to erroneous correlations within the model. Facial images in the elderly category exhibit features closely associated with aging, such as nasolabial folds, wrinkled skin, and subtler expressions. Consequently, fairness metrics in this age group are inferior compared to other age groups. Despite this, the proposed method succeeds in enhancing fairness metrics.

D. Ablation Study

Table VI reports the results of the ablation study conducted across three categories. Initially, we explore bias mitigation in attention, targeting the query, key, value, and all the three combined. Regularizing either the query or key, both associated with landmarks, results in a decrease of accuracy. In contrast, regularizing the value successfully mitigates age bias and improves accuracy. Regularizing all of them yields intermediate results, closer to those of value regularization.

TABLE VI. RESULTS ON ABLATION STUDY

Category	Type	Acc.	EOP	EOD
Baseline		0.9219	0.3227	0.1694
Triplet target	Q-K	0.9313	0.2612	0.1372
	Value	0.9237	0.3280	0.1721
Mitigation target	Q-K	0.9217	0.3187	0.1683
	Value	0.9313	0.2612	0.1372
Latent vector dimension	32	0.9207	0.2549	0.1674
	64	0.9313	0.2612	0.1372
	128	0.9290	0.2591	0.1381
	256	0.9292	0.2611	0.1352
Absence of module	D_a	0.9117	0.3241	0.1702
	C_e	0.9107	0.2603	0.1427

The second part of the study alters the dimension of the latent vector. Both accuracy and bias mitigation metrics tend to increase as the latent vector's dimension increases. This suggests that higher dimensions allow for smoother bias mitigation and information preservation. However, beyond a dimension of 64, no substantial difference is observed, indicating 64 as the optimal dimension for this task.

The third aspect examines the impact of removing modules constituting the autoencoder. Removing the age discriminator leads to the results similar to the baseline, indicating no regularization. In the absence of the emotion classifier, bias is further reduced, but accuracy substantially decreases.

V. CONCLUDING REMARKS

In this paper, we have proposed a facial expression recognition (FER) model that mitigates age-related bias by decoupling and independently regularizing the value vectors and query-key matrices within the self-attention mechanism of Transformer. Based on the observation that value vectors encode age-related facial textures while query-key matrices reflect landmark-based emotion cues, tailored regularization strategies have been applied to each component. An adversarial autoencoder is used to remove age-related information from the value vectors while preserving emotional semantics. Triplet learning is applied to the query-key matrices to address data imbalance and enhance the separation of emotional representations.

Experimental results demonstrate that the proposed method achieves superior performance compared to existing bias mitigation techniques. It notably reduces age-related disparities in classification outcomes while maintaining high overall accuracy. Furthermore, it successfully alleviates negative prediction tendencies toward older individuals and neutral expressions without degrading predictive capability, showing promise for fairer and more robust FER systems.

In future work, we consider the following directions. First, we aim to extend our framework to address multi-attribute bias mitigation, targeting other social attributes such as gender and race in addition to age. Since facial images often contain entangled demographic features, it is essential to develop disentanglement and regularization techniques that mitigate bias for multiple attributes simultaneously without compromising fairness for any single factor. Second, to support real-time FER applications, future research will focus on optimizing the proposed model for efficiency and lightweight deployment. Given that the proposed regularization operates at the self-attention layer level, a deeper understanding of Transformer architecture is required to strategically place and scale the regularization components for maximum performance and minimal overhead. Ultimately, this line of research can lead to facial recognition systems that are not only accurate but also fair and inclusive, contributing to the development of trustworthy AI technologies and a more equitable society.

CREDiT AUTHORSHIP CONTRIBUTION STATEMENT

Jaecil Park: software, validation, investigation, visualization, methodology, writing – original draft.

Sung-Bae Cho: conceptualization, methodology, supervision, funding acquisition, project administration, writing – review and editing.

DATA STATEMENT

The study used publicly available datasets.

DECLARATION OF CONFLICTS OF INTEREST

We have no conflict of interest to declare.

ACKNOWLEDGMENT

This work was supported by IITP grant funded by the Korean government (MSIT) (No. RS-2022-II220113, Developing a Sustainable Collaborative Multi-modal Lifelong Learning Framework), and Air Force Defense Research Sciences Program funded by Air Force Office of Scientific Research.

REFERENCES

- [1] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: development and applications to human computer interaction," in *2003 Conference on computer vision and pattern recognition workshop*, vol. 5, pp. 53-53, 2003.
- [2] Z. Fei, E. Yang, D. D.-U. Li, S. Butler, W. Ijomah, X. Li, and H. Zhou, "Deep convolution network based emotion analysis towards mental health care," *Neurocomputing*, vol. 388, pp. 212-227, 2020.
- [3] J. Deng, G. Pang, Z. Zhang, Z. Pang, H. Yang, and G. Yang, "cGAN based facial expression recognition for human-robot interaction," *IEEE Access*, vol. 7, pp. 9848-9859, 2019.
- [4] F. Xue, Q. Wang, Z. Tan, Z. Ma, and G. Guo, "Vision transformer with attentive pooling for robust facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3244-3256, 2022.
- [5] S. Biswas and J. Sil, "An efficient expression recognition method using contourlet transform," in *International Conference on Perception and Machine Intelligence*, pp. 167-174, 2015.
- [6] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski, "A survey on facial emotion recognition techniques: A state-of-the-art literature review," *Information Sciences*, vol. 582, pp. 593-617, 2022.
- [7] H. Wang, V. Sanchez, and C.-T. Li, "Improving face-based age estimation

- with attention-based dynamic patch fusion,” *IEEE Transactions on Image Processing*, vol. 31, pp. 1084-1096, 2022.
- [8] R. Angulu, J. R. Tapamo, and A. O. Adewumi, “Age estimation via face images: a survey,” *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, pp. 1-35, 2018.
 - [9] T. Ganel, C. Sofer, and M. A. Goodale, “Biases in human perception of facial age are present and more exaggerated in current AI technology,” *Scientific Reports*, vol. 12, no. 1, p. 22519, 2022.
 - [10] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, and A. Huang, “POSTER++: A simpler and stronger facial expression recognition network,” *arXiv preprint arXiv:2301.12149*, 2023.
 - [11] H. Li, M. Sui, F. Zhao, Z. Zha, and F. Wu, “MVT: mask vision transformer for facial expression recognition in the wild,” *arXiv preprint arXiv:2106.04520*, 2021.
 - [12] G. S. Hayes, S. N. McLennan, J. D. Henry, L. H. Phillips, G. Terrett, P. G. Rendell, R. M. Pelly, and I. Labuschagne, “Task characteristics influence facial emotion recognition age-effects: A meta-analytic review,” *Psychology and aging*, vol. 35, no. 2, p. 295, 2020.
 - [13] Y. Huang, F. Chen, S. Lv, and X. Wang, “Facial expression recognition: A survey,” *Symmetry*, vol. 11, no. 10, p. 1189, 2019.
 - [14] T. Xu, J. White, S. Kalkan, and H. Gunes, “Investigating bias and fairness in facial expression recognition,” in *European Conference on Computer Vision Workshops*, pp. 506-523, 2020.
 - [15] Y. Fan, J. C. Lam, and V. O. Li, “Unsupervised domain adaptation with generative adversarial networks for facial emotion recognition,” in *IEEE International Conference on Big Data*, pp. 4460-4464, 2018.
 - [16] A. H. Sham, K. Aktas, D. Rizhinashvili, D. Kuklianov, F. Alisanoglu, I. Ofodile, C. Ozcinar, and G. Anbarjafari, “Ethical AI in facial expression analysis: racial bias,” *Signal, Image and Video Processing*, vol. 17, no. 2, pp. 399-406, 2023.
 - [17] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, “Joint fine-tuning in deep neural networks for facial expression recognition,” in *IEEE International Conference on Computer Vision*, pp. 2983-2991, 2015.
 - [18] S. Li, W. Deng, and J. Du, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2852-2861, 2017.
 - [19] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “Affectnet: A database for facial expression, valence, and arousal computing in the wild,” *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18-31, 2017.
 - [20] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, and D.-H. Lee, “Challenges in representation learning: A report on three machine learning contests,” in *International Conference on Neural Information Processing*, pp. 117-124, 2013.
 - [21] J. Shi, S. Zhu, and Z. Liang, “Learning to amend facial expression representation via de-albino and affinity,” *arXiv preprint arXiv:2103.10189*, 2021.
 - [22] L. Pham, T. H. Vu, and T. A. Tran, “Facial expression recognition using residual masking network,” in *International Conference on Pattern Recognition*, pp. 4513-4519, 2021.
 - [23] Y. Chen, J. Li, S. Shan, M. Wang, and R. Hong, “From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos,” *IEEE Transactions on Affective Computing*, 2024.
 - [24] C. Wang, L. Chen, L. Wang, Z. Li, and X. Lv, “QCS: Feature refining from quadruplet cross similarity for facial expression recognition,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 7, pp. 7563-7572, 2025.
 - [25] G. Guo, R. Guo, and X. Li, “Facial expression recognition influenced by human aging,” *IEEE Transactions on Affective Computing*, vol. 4, no. 3, pp. 291-298, 2013.
 - [26] E. Kim, D. A. Bryant, D. Srikanth, and A. Howard, “Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle-aged, and older adults,” in *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 638-644, 2021.
 - [27] A. Howard, C. Zhang, and E. Horvitz, “Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems,” in *2017 IEEE Workshop on Advanced Robotics and its Social Impacts*, pp. 1-7, 2017.
 - [28] B. H. Zhang, B. Lemoine, and M. Mitchell, “Mitigating unwanted biases with adversarial learning,” in *2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335-340, 2018.
 - [29] J. Cai, Z. Meng, A. S. Khan, J. O’Reilly, Z. Li, S. Han, and Y. Tong, “Identity-free facial expression recognition using conditional generative adversarial network,” in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 1344-1348, 2021.
 - [30] M. Ren, W. Zeng, B. Yang, and R. Urtasun, “Learning to reweight examples for robust deep learning,” in *International conference on Machine Learning*, pp. 4334-4343, 2018.
 - [31] M. Ochal, M. Patacchiola, J. Vazquez, A. Storkey, and S. Wang, “Few-shot learning with class imbalance,” *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 5, pp. 1348-1358, 2023.
 - [32] R. A. Bauder and T. M. Khoshgoftaar, “The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data,” *Health information science and systems*, vol. 6, pp. 1-14, 2018.
 - [33] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, “Data imbalance in classification: Experimental evaluation,” *Information Sciences*, vol. 513, pp. 429-441, 2020.
 - [34] F. Hoitsma, G. Nápoles, Ç. Güven, and Y. Salgueiro, “Mitigating implicit and explicit bias in structured data without sacrificing accuracy in pattern classification,” *AI & SOCIETY*, pp. 1-20, 2024.
 - [35] J. Ren, C. Yu, X. Ma, H. Zhao, and S. Yi, “Balanced meta-softmax for long-tailed visual recognition,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 4175-4186, 2020.
 - [36] C. Wei, K. Sohn, C. Mellina, A. Yuille, and F. Yang, “Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10857-10866, 2021.
 - [37] A. T. Wasi, T. H. Rafi, R. Islam, K. Šerbetar, and D.-K. Chae, “GrEFEL: Geometry-Aware Reliable Facial Expression Learning under Bias and Imbalanced Data Distribution,” in *Proceedings of the Asian Conference on Computer Vision*, pp. 4368-4384, 2024.
 - [38] M. Sreenivas, S. Takamuku, S. Biswas, A. Chepuri, B. Vengatesan, and N. Natori, “Improved Cross-Dataset Facial Expression Recognition by Handling Data Imbalance and Feature Confusion,” in *European Conference on Computer Vision*, pp. 262-277, 2022.
 - [39] A. Vaswani, “Attention is all you need,” *arXiv preprint arXiv:1706.03762*, 2017.
 - [40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *International Conference on Learning Representations*, 2020.
 - [41] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, “Transformers in vision: A survey,” *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1-41, 2022.
 - [42] H. Li, Z. Guo, S.-M. Rhee, S. Han, and J.-J. Han, “Towards accurate facial landmark detection via cascaded transformers,” in *IEEE/CVF Conference on Vision and Pattern Recognition*, pp. 4176-4185, 2022.
 - [43] Y. Tewel, R. Gal, G. Chechik, and Y. Atzmon, “Key-locked rank one editing for text-to-image personalization,” in *ACM SIGGRAPH 2023 Conf. Proceedings*, pp. 1-11, 2023.
 - [44] Z. Li, D. Gong, X. Li, and D. Tao, “Aging face recognition: a hierarchical learning model based on local patterns selection,” *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2146-2154, 2016.
 - [45] J. Park and S.-B. Cho, “Age-unbiased facial emotion recognition with regularizing self-attention value vector,” in *International Conference on Intelligent Data Engineering and Automated Learning*, pp. 472-480, 2024.
 - [46] J.-Y. Kim and S.-B. Cho, “An information theoretic approach to reducing algorithmic bias for machine learning,” *Neurocomputing*, vol. 500, pp. 26-38, 2022.
 - [47] H.-J. Moon, S.-J. Bu, and S.-B. Cho, “Learning disentangled representation of residential power demand peak via convolutional-recurrent triplet network,” in *International Conference on Data Mining Workshops*, pp. 757-761, 2020.
 - [48] J. Park and S.-B. Cho, “Adversarial Discriminator to Mitigate Gender Bias in Abusive Language Detection,” in *European Conference on Artificial Intelligence (ECAI 2023)*, Frontiers in Artificial Intelligence and Applications, vol. 372, pp. 1851-1858, 2023.
 - [49] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, “Adversarial autoencoders,” *arXiv preprint arXiv:1511.05644*, 2015.
 - [50] I. Chen, F. D. Johansson, and D. Sontag, “Why is my classifier

discriminatory?,” in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

- [51] J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon, “Learning controllable fair representations,” in *International Conference on Artificial Intelligence and Statistics*, pp. 2164-2173, 2019.
- [52] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” *arXiv preprint arXiv:1703.07737*, 2017.
- [53] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815-823, 2015.
- [54] N. C. Ebner, M. Riediger, and U. Lindenberger, “FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation,” *Behavior Research Methods*, vol. 42, pp. 351-362, 2010.
- [55] T. Ruffman, J. D. Henry, V. Livingstone, and L. H. Phillips, “A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging,” *Neuroscience & Biobehavioral Reviews*, vol. 32, no. 4, pp. 863-881, 2008.
- [56] S. G. Müller and F. Hutter, “Trivialaugment: Tuning-free yet state-of-the-art data augmentation,” in *IEEE/CVF International Conference on Computer Vision*, pp. 774-782, 2021.
- [57] R. Takahashi, T. Matsubara, and K. Uehara, “Data augmentation using random image cropping and patching for deep CNNs,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 2917-2931, 2019.
- [58] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” in *Advances in Neural Information Processing Systems*, vol. 29, 2016.



Jaeil Park

Jaeil Park received the B.S. degree in computer science from Yonsei University, Seoul, South Korea, where he is currently pursuing the integrated M.S. and Ph.D. program in the Department of Computer Science. His research interests include artificial intelligence fairness, explainability, and trustworthy machine learning systems.



Sung-Bae Cho

Sung-Bae Cho received the M.S. and Ph.D. degrees in computer science from KAIST (Korea Advanced Institute of Science and Technology), Korea. He was an invited researcher of Human Information Processing research laboratories at ATR (Advanced Telecommunications Research) institute, Japan from 1993 to 1995, and a visiting scholar at University of New South Wales, Australia in 1998.

He was also a visiting professor at University of British Columbia, Canada from 2005 to 2006, and at King Mongkut's University of Technology at Thonburi, Thailand in 2013. Since 1995, he has been a professor in the Department of Computer Science, Yonsei University, a Underwood distinguished professor from 2021, and a Yonsei Lee Youn-Jae Fellow from 2023. His research interests include neural networks, pattern recognition, and intelligent man-machine interfaces. He was the recipient of the Richard E. Merwin prize from the IEEE Computer Society in 1993. He received several awards from Korea Information Science Society in 2005, and Gaheon Sindoricoh in 2017. He is also a recipient of Service Merit Medal from Korean government in 2022, and the Sudang prize in 2026. Currently he is the Fellow of IEEE, Korean Academy of Science and Technology, and National Academy of Engineering of Korea.