IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 February 2024, accepted 3 May 2024, date of publication 13 May 2024, date of current version 20 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3400150

RESEARCH ARTICLE

Toward Identity-Invariant Facial Expression Recognition: Disentangled Representation via Mutual Information Perspective

DAEHA KIM, SEONGHO KIM, (Associate Member, IEEE), AND BYUNG CHEOL SONG, (Senior Member, IEEE)
Department of Electrical and Computer Engineering, Inha University, Incheon 22212, Republic of Korea

Corresponding author: Byung Cheol Song (bcsong@inha.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by Korean Government (MSIT), Artificial Intelligence Convergence Innovation Human Resources Development, Inha University, under Grant RS-2022-00155915 and Grant IITP-2021-0-02052; and in part by the National Research Foundation of Korea (NRF) Grant funded by Korean Government (MSIT) under Grant 2022R1A2C2010095 and Grant 2022R1A4A1033549.

ABSTRACT This paper presents an identity-invariant facial expression recognition framework. It aims to make a facial expression recognition (FER) model independently understand facial expressions and identity (ID) attributes such as gender, age, and skin, which are entangled in face images. The learned representations of the FER model pursue robustness against unseen ID samples with large attribute differences. Specifically, attribute properties describing (facial) images are retrieved through a powerful pre-trained model, i.e., CLIP. Then, expression features and ID features are realized through residual module(s). As a result, the features learn expression-efficient and ID-invariant representations based on mutual information. The proposed framework is compatible with various backbones, and enables detachment/attachment of ID attributes and ablative analysis. Extensive experiments for several wild Valence-Arousal domain databsets showed the performance improvement of maximum 9% compared to the runner up, and also demonstrated the subjective realism of ID-invariant representation in high-dimensional image space.

INDEX TERMS CLIP, demographics, facial expression, facial identity, mutual information.

I. INTRODUCTION

Facial expression recognition (FER) technology plays an important role in Human-Computer Interaction (HCI) because it can instantly recognize a person's emotional state with only facial expression. Unlike so-called multi-modal emotion recognition tasks [1], [2] that perform emotion recognition through various modalities (e.g., speech and EEG), facial expression recognition is classified as a challenging task due to the sole reliance on facial expressions to recognize emotions that can be subjectively interpreted. In such challenging conditions, the emotion domain that FER needs to predict is segmented into discrete and continuous domains. Specifically, the emotional state can be expressed by a discrete domain model [3] in which only a finite number of emotions are annotated, or a continuous domain model [4] based on two axes of valence (V) and arousal (A). Here, V and A indicate the degree of positive/negative emotions and the intensity of activation, respectively. With the advent of large-scale VA datasets [5], [6], VA FER methods for mapping facial images to VA space have already achieved reliable performance enough to be used in the real world [7], [8].

On the other hand, in order to get the generalized performance, it must be robust against unseen identity (ID) because the expression tendency of model training is generally different from that of unseen ID. This task, named as ID-invariant FER ($\mathrm{I^2FER}$), has been seldom tackled due to two challenging points: 1) In face images, not only ID attributes (e.g., gender, age, skin) but also semantic factors such as facial expressions are entangled, so different expressions of the same ID can be clustered into one. 2) Facial

The associate editor coordinating the review of this manuscript and approving it for publication was Binit Lukose.

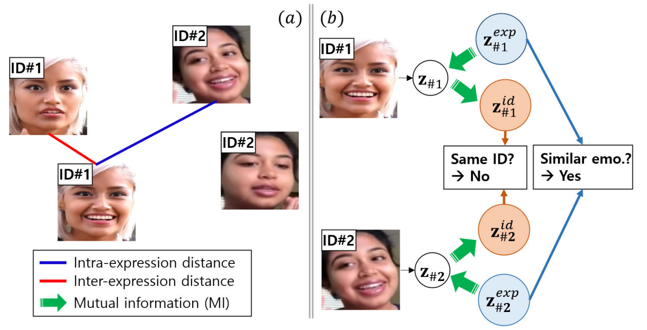


FIGURE 1. ***(a)* t-SNE [9] visualization from a SOTA FER model [8]. Since this model interprets features dependent on IDs rather than expressions themselves, similar emotions of ID#1 and ID#2 may be recognized differently. *(b)* Concept of our idea. ID#1 and ID#2 are judged to have the same emotion thanks to MI.**

features should be informative about expressions and robust to changes in IDs because the same emotions can be often recognized differently due to ID attributes.

Figure 1(*a*) shows a typical ID dependency case where identical IDs are closer to each other than similar expressions. In general, since facial images include both ID-dependent and expression-dependent characteristics [10], variations in ID attributes are accompanied by changes in expression.

A few studies [11] emerged, noting the importance of observing as many emotions as possible to properly tackle I^{2}FER. For example, it was shown that grouping by emotion intensity [12] or learning the expression consistency of different IDs [13] provided an opportunity to observe various expression tendencies of IDs. Recently, [8] captured the optimal expression tendency between IDs through optimal transport theory. Although this study is encouraging in terms of dealing with ID dependency for the first time, ID dependency phenomenon is still observed (cf. Sec. V-B). In other words, [8] does not structurally separate ID-dependent components, which is an obstacle to I^{2}FER. More generally, InvarPS [14] and IBN-Net [15] were developed asserting the necessity of separating unnecessary features entangled with the target domain in several vision domains. Their arguments are somewhat consistent with the fact that the ID-invariant feature issue observed in Fig. 1 is actually crucial in the face image domain.

Therefore, explicit separation of ID-dependent features from facial features will be the key to solving I^{2}FER. We propose a novel EIF framework that learns expression-efficient representations using Expression feature z^{exp}, ID feature z^{id}, and Facial feature z, which are independent each other. Orthogonal to other VA FER methods, the EIF framework employs demographics representing a specific group of subjects as self-supervision, along with facial images. Also, a powerful pre-trained model, i.e., CLIP [16], provides attribute features z^{att} suitable for a given facial image with several demographics as a questionnaire. Then, z^{id} and z^{exp} are generated from z^{att} and z, respectively (cf. Sec. IV-B). Finally, we design a novel mechanism that simultaneously learns ID attributes and expressions from z^{id} and z^{exp} of different attributes (see Fig. 1(*b*)). The proposed method is based on mutual information (MI) and is realized through an objective of making z informative with z^{exp} but invariant with z^{id} through disentanglement of z^{id} and z^{exp}. Thus, the model trained with this objective is expected to be able to predict expressions robustly even in the ID attributes of unseen IDs (cf. Sec. III-B).

Our main contributions are summarized as follows:

- We issue on a challenging ID-invariant FER (I^{2}FER) by exploring ID attributes that are key to FER. It is the first case in the FER field that tackles the generalization of FER by encoding ID attributes into high-dimensional features.
- A novel input combination of facial images and demographics, which has not been tried before, provides ID (and expression)-dependent components as features. In particular, the MI-induced objective disentangles the two features to enable learning of expression-dependent representations.
- We demonstrate the validity of EIF framework through extensive simulations on wild VA datasets. Especically, the feature-extrapolated downstream task qualitatively verifies the ID-invariant representations of EIF in image space (cf. Fig. 9).

II. RELATED WORK

A. VA FER AND BEYOND

VA FER, which annotates the intensity and degree of emotion on the circular space of V and A, ultimately pursues FER in a wild environment including complex and micro-emotions. Recently, thanks to large-scale VA databases [5], [6], [17] and convolutional neural networks (CNNs) [18], [19], facial expression correlations between facial images and labels have been successfully formulated. For example, the spectrum of related research is expanding from a CNN model based on residual connection [20] to feature learning through adversarial auto-encoding [21].

Towards **I^{2}FER***:* For the real-world application of FER technology, the ability to cover unseen IDs is essential. For example, Ali and Hughes [13] tried I^{2}FER by learning expression consistency from similar expressions between different IDs. Also, several studies have been published to achieve generalization of FER by learning the emotional diversity [7], [12]. Recently, Kim and Song [8] quantified ID shifts reflecting the expression tendency of samples through optimal transport theory, and then learned ID-invariant representations from ID shifts. However, FER approach so far has not been able to explicitly separate ID (or expression)-dependent components from features due to the absence of a powerful encoder or theoretical objective design. This problem causes the ID dependency phenomenon in which ID attributes are involved in expression changes in image space (cf. Fig. 9).

B. DISENTANGLED REPRESENTATION FROM AN INFORMATION THEORY PERSPECTIVE

The goal of disentangled representation learning is to define or generate explanatory factors in a latent space. Feature

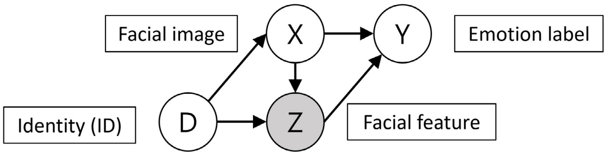


FIGURE 2. Illustration of causal diagram for FER, given four variables, i.e., facial image X, identity D, facial feature Z, and emotion label Y. Direct edge of this causal diagram represents the causality between the two nodes, i.e., the cause → effect. Here, white and gray represent observed and latent features, respectively.

disentanglement for this learning scheme is mainly done inside the model through the parametric module or in the loss stage. The latter approach figured out many vision tasks through various metrics (e.g., Wasserstein, mean discrepancy etc.) [22], [23], [24], [25]. In particular, MI [26] is notable because it is suitable for handling high-dimensional features and has already been verified in high-level tasks [27], [28], [29]. For example, Chen et al. [28] disentangled features with different properties by maximizing MI between latent features and data. Savarese et al. [29] proposed a segmentation task of in-painted foreground and background through this MI-based model learning.

Furthermore, representation learning based on the theory of information bottleneck [30] has recently been attracting attention [31], [32], [33], [34], [35], [36]. Specifically, the MI-induced objective is designed so that the learning scheme is effective for variables useful for the task and is invariant to obstacle variables. Inspired by the learning scheme, we define a complementary objective that maximizes MI between facial features and expression features while minimizing MI between ID features, and learns expression-efficient representations.

III. PRELIMINARIES

This section explores the I^{2}FER using the causal diagram as a tool, and designs an ID-invariant objective for disentanglement of ID (and expression)-dependent components.

A. PROBLEM FORMULATION

This paper aims at VA FER, a regression task that estimates VA label $\mathbf{y}^{\mathrm{va}}(\in Y)$ from (facial) feature $\mathbf{z}(\in Z)$. In $\mathbf{z}$ where ID (and expression)-dependent components are mixed, features of the same ID tend to get closer in the feature space [37]. As a result, this causes a bottleneck in FER performance for the test-set containing many unseen IDs. To observe at which stage of the (VA) FER ID dependency phenomenon occurs, we introduce a (structural) causal diagram [38], [39]. Causal diagram defines cause and effect, that is, causality, with each variable as a node. Using the causal diagram, we analyze which variable is an ID-dependent component by regarding the ID attribute as (initial) cause and the emotion label as (final) effect, respectively. Especially, since the relationship between all variables can be examined one by one through this diagram, it is useful for exploring the FER task in which various attributes are combined with facial features.

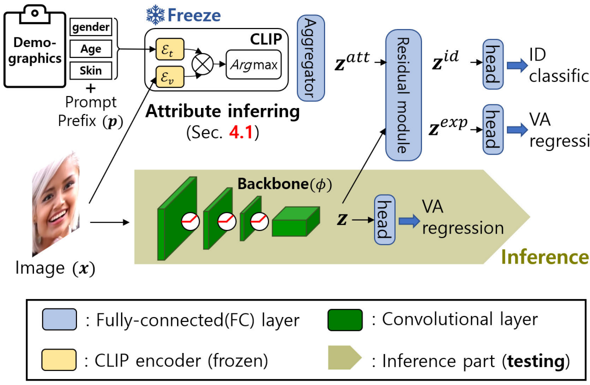


FIGURE 3. An overview of EIF framework. Two features obtained through backbone (ϕ) and CLIP are factorized into ID and expression features respectively via residual module. ϕ learns an ID-invariant representation by different types of losses. In other words, z^{id} and z^{exp} only provide θ with the opportunity to learn ID-invariant features, and have no separate role in the testing phase. Therefore, inference of emotion is performed only with z.

As in Fig. 2, there are three causalities associated with the facial feature Z: $X \rightarrow Z, Z \rightarrow Y$, and $D \rightarrow Z$. First, $X \rightarrow Z$ and $Z \rightarrow Y$ are cases where the effect always changes as the cause changes. From the FER point of view, a change in facial expression causes a change in emotion label, which is because facial features must also change. However, $D \rightarrow Z$ is not a case in which the effect must always change according to cause's change. This is because each subject contains its own expression. This fact requires the following prerequisite.

Prerequisite 1: Z should be informative to Y and independent of D as much as possible.

Using the MI-induced objective in the upcoming section, we present a clear solution to this prerequisite. The tendency of D to express emotions transfers to ID-specific facial expressions in image X (see $D \rightarrow X$ in Fig. 2). It's important to note that this paper focuses solely on variables directly associated with feature Z.

B. ID-INVARIANT OBJECTIVE

Mutual information has been frequently used to learn dependencies between task-efficient variables in domain adaptation or generalization fields [40], [41], [42]. In general, minimization of MI in the feature space encourages model learning to be invariant to target variables. With this in mind, we design the objective so that Z is informative to Y, but is invariant to D.

$$\max_{Z} \mathrm{MI}(Z, Y) - \mathrm{MI}(Z, D). \tag{1}$$

where using Z as an anchor, the relationship between Y and D is simultaneously learned, and this formulation coincides with the causal diagram of Fig. 2. Specifically, the maximization of $\mathrm{MI}(Z, Y)$ and the minimization of $\mathrm{MI}(Z, D)$ encourage the model to learn expression-efficient and ID-invariant representations, respectively. Furthermore, when the parametric model ϕ encodes X into Z, the above

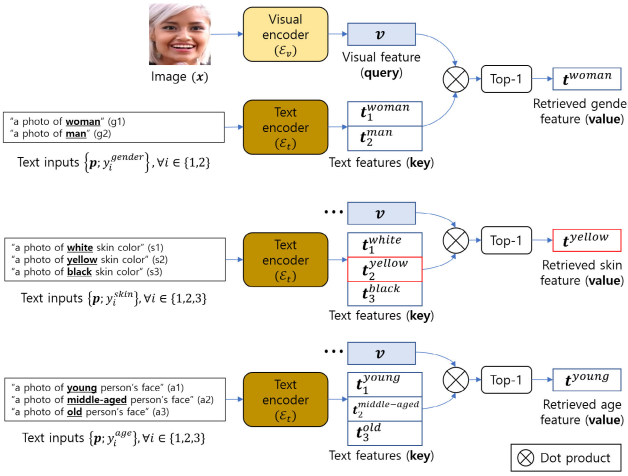


Image (x)
Visual encoder ($\mathcal{E}_v$)
v
Visual feature (query)
Top-1
t^{woman}
Retrieved gender feature (value)
"a photo of woman" (g1)
"a photo of man" (g2)
Text inputs $\{p; y_i^{gender}\}, \forall i \in \{1,2\}$
Text encoder ($\mathcal{E}_t$)
t_1^{woman}
t_2^{man}
Text features (key)
v
t_1^{white}
t_2^{yellow}
t_3^{black}
Top-1
t^{yellow}
Retrieved skin feature (value)
"a photo of white skin color" (s1)
"a photo of yellow skin color" (s2)
"a photo of black skin color" (s3)
Text inputs $\{p; y_i^{skin}\}, \forall i \in \{1,2,3\}$
Text encoder ($\mathcal{E}_t$)
Text features (key)
v
t_1^{young}
$t_2^{middle-aged}$
t_3^{old}
Top-1
t^{young}
Retrieved age feature (value)
"a photo of young person's face" (a1)
"a photo of middle-aged person's face" (a2)
"a photo of old person's face" (a3)
Text inputs $\{p; y_i^{age}\}, \forall i \in \{1,2,3\}$
Text encoder ($\mathcal{E}_t$)
Text features (key)
Dot product

FIGURE 4. Pipeline of CLIP encoding with demographic inputs. The retrieved three vectors t^{woman}, t^{yellow}, and t^{young} are used to get an attribute vector z^{att} describing an image x.

objective is re-written by

$$\max_{\phi} \mathrm{MI}(\phi(X), Y) - \mathrm{MI}(Y, D|\phi(X)). \tag{2}$$

Reason for deriving the second term of Eq. 2 is as follows.

Lemma 1: Minimizing $\mathrm{MI}(Y, D|\phi(X))$ is equivalent to learning ϕ so that $\phi(X)$ and D disentangle from each other.

Proof: According to the property of conditional MI, $\mathrm{MI}(Y, D|\phi(X)) = \mathrm{MI}(Y, D, \phi(X)) - \mathrm{MI}(Y, \phi(X))$. *As* $\mathrm{MI}(Y, D|\phi(X))$ *is minimized,* $\mathrm{MI}(Y, D, \phi(X)) = \mathrm{MI}(Y, \phi(X))$. *Therefore, D is invariant with $\phi(X)$.* □

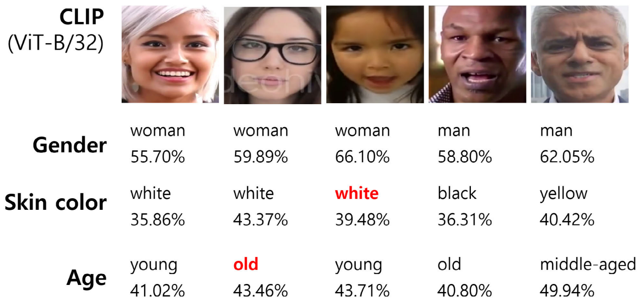
IV. PROPOSED METHOD: EIF

To implement the ID-invariant objective defined in the previous section, this section focuses on two points: Input features of the objective and formulation of the objective in a tractable form. Due to the unique structure of the EIF framework reviewed in Fig. 3, images and demographics are input together. Then, let's take a look at how to manipulate the inputs.

Visual Input: An image $\boldsymbol{x}(\in X)$ is encoded into a facial feature $z(\in Z)$ through the backbone (ϕ). In parallel, the ID label y^{id} corresponding to $\boldsymbol{x}$ is given from the ID index assigned to each image sample (cf. Sec. V for details).

Demographic: To retrieve ID attributes matching $\boldsymbol{x}$, we build a demographic set $\mathcal{D} = \{\text{skin, age, gender}\}$ using well-known face-related taxonomy [43] (i.e., skin color, age, gender). For example, three texts of 'yellow', 'white', and 'black' are provided for the skin attribute label y_i^{skin}: $\forall i \in \{1, 2, 3\}$. Then, attribute-specific text inputs $\{\boldsymbol{p}; y_i^{\text{skin}}\}$ from the hand-crafted prompt prefix, i.e., $\boldsymbol{p}$ ='a photo of a ' and y_i^{skin} are defined. Also, specific text inputs corresponding to age and gender are provided. The next section describes the process of encoding z^{att} from the self-supervision inputs generated from $\mathcal{D}$.

The Rationale for Replacing Race With Skin Color: The 'race' attribute, introduced in Fairface [43], intuitively groups subjects. However, since race is an abstract concept that is difficult to represent only with visual information, even CLIP [16] is hard to encode this attribute. So, we replace race with 'skin color' which indirectly reflects it.



CLIP (ViT-B/32)					
Gender	woman 55.70%	woman 59.89%	woman 66.10%	man 58.80%	man 62.05%
Skin color	white 35.86%	white 43.37%	white 39.48%	black 36.31%	yellow 40.42%
Age	young 41.02%	old 43.46%	young 43.71%	old 40.80%	middle-aged 49.94%

FIGURE 5. Fidelity of CLIP encoded with demographic set. Red text indicates incorrect inferring results. In the case of the image in the second column, it looks as if the age attribute was inferred as 'old' because the artifacts act as wrinkles on the face.

Details of CLIP Encoding: To get the attribute feature (z^{att}), we employ CLIP having powerful zero-shot encoding capability. The detailed process is shown in figure 4. First, we feed facial image ($\boldsymbol{x}$) and demographic set ($\mathcal{D}$ = {skin, age, gender}) to the visual encoder ($\mathcal{E}_v$) and text encoder ($\mathcal{E}_t$) of CLIP, respectively. Then, visual features ($\boldsymbol{v}$) and text features ($\boldsymbol{t}_i^{\mathcal{D}}$) are output, respectively. Next, with $\boldsymbol{v}$ as a query feature and $\boldsymbol{t}_i^{\mathcal{D}}$ as key features, text features that is the key to describe $\boldsymbol{v}$ are retrieved through 'Top-1' operation. In detail, this is implemented by comparing dot products between the two features and selecting the maximum. When the similarity between $\boldsymbol{v}$ and $\boldsymbol{t}_2^{\text{skin}}$ is the largest, $\boldsymbol{t}_2^{\text{skin}}$ is retrieved as a feature representing the skin property of $\boldsymbol{x}$ (see the red box in Fig. 4). This process is performed for all elements of $\mathcal{D}$. In the end, $\boldsymbol{v}$ and the retrieved text features are concatenated and used as input to the FC layer.

Fidelity of CLIP Encoding: Since the proposed method relies on CLIP to generate z^{att}, it is essential to check the validity of CLIP for facial images. So, we show that CLIP is actually effective in the face domain by presenting the attribute inferring results for some samples. The inferring results through CLIP (ViT-B/32) is given in Fig.5. In the case of the gender attribute that distinguishes woman from man, human-level predictions were shown. However, in the case of age, some unexpected errors were found (see the second column in Fig. 5). Fortunately, such results occasionally occurred when artifacts were mixed in the image. Also, since the results for other attributes come out as we intended, these outlier cases are relatively negligible. In the future, if a face attribute-aware encoding module is developed, more sophisticated z^{att} can be generated.

A. ATTRIBUTE INFERRING

Since previous FER models [12], [20] never considered demographic(s), they could not explicitly use ID attributes for FER learning. Note that ID attributes provide an opportunity to focus only on expression-efficient representations while keeping $\boldsymbol{z}$ invariant to ID. So, this section presents a novel attribute inferring that generates $\boldsymbol{z}^{\text{att}}$ suitable for $\boldsymbol{x}$.

1) FEATURE ENCODING

CLIP [16] has recently been spotlighted as a model that learns contrastive loss so that two feature spaces are aligned from a pair of visual and text input. CLIP encoders $\mathcal{E}_v, \mathcal{E}_t$ encode visual and text inputs into the visual feature $\boldsymbol{v}$ and the text features $\boldsymbol{t}_i^{\mathcal{D}}$, respectively: $\boldsymbol{v} = \mathcal{E}_v(\boldsymbol{x})$ and $\boldsymbol{t}_i^{\mathcal{D}} = \mathcal{E}_t(\{\boldsymbol{p}; y_i^{\mathcal{D}}\})$.

2) ATTRIBUTE INFERRING

Next, to infer $\boldsymbol{t}_i^{\mathcal{D}}$ that best describes $\boldsymbol{x}$, we calculate the prediction probability of $\boldsymbol{v}$ through the cosine similarity (cos) of $\boldsymbol{v}$ and $\boldsymbol{t}_i^{\mathcal{D}}$ as follows:

$$p(y_i^{\mathcal{D}}|\boldsymbol{x}) = \frac{exp(\cos(\boldsymbol{v}, \boldsymbol{t}_i^{\mathcal{D}})/\tau)}{\sum_{i'} exp(\cos(\boldsymbol{v}, \boldsymbol{t}_{i'}^{\mathcal{D}})/\tau)}, \qquad i^* := \arg\max_i p(y_i^{\mathcal{D}}|\boldsymbol{x}), \tag{3}$$

where τ indicates the temperature of the softmax. From the perspective of skin, age, and gender, $\boldsymbol{t}_{i^*}^{\mathcal{D}}$s describing $\boldsymbol{x}$ are retrieved, respectively. In detail, $\boldsymbol{v}$ and $\boldsymbol{t}_{i^*}^{\mathcal{D}}$ are concatenated, and passes through a single fully-connected (FC) layer, and is encoded into $\boldsymbol{z}^{\text{att}}$: $\boldsymbol{z}^{\text{att}} := \text{FC}([\boldsymbol{v}; \boldsymbol{t}_{i^*}^{\mathcal{D}}])$.[1] Section V-E verifies the validity of $\boldsymbol{z}^{\text{att}}$ as well as the number of attributes.

B. RESIDUAL MODULE

To learn emotion patterns and ID-invariant representations at the same time, this section describes how to generate $\boldsymbol{z}^{\text{exp}}$s and $\boldsymbol{z}^{\text{id}}$s. Inspired by the fact that face images contain both expression and attribute features [10], $\boldsymbol{z}$ can be defined by $\boldsymbol{z} = \boldsymbol{z}^{\text{exp}} + \boldsymbol{z}^{\text{att}}$. Based on this relationship, we propose a residual module $\mathcal{R}$ that encodes ID (and expression)-dependent components into features as follows:

$$\boldsymbol{z}^{\text{exp}} = \mathcal{R}(\boldsymbol{z} - \boldsymbol{z}^{\text{att}}) \quad \text{and} \quad \boldsymbol{z}^{\text{id}} = \mathcal{R}(\boldsymbol{z}^{\text{att}}), \tag{4}$$

where $\mathcal{R}$ is designed as a single FC layer with leaky ReLU. Weight-shared $\mathcal{R}$ has two goals: 1) Generation of $\boldsymbol{z}^{\text{exp}}$ from feature residual components and 2) generation of $\boldsymbol{z}^{\text{id}}$ through the refinement of attributes.

C. OBJECTIVE FORMULATION

From the generated features and Eq. 2, the objective for ID-invariant expression-efficient representations is formulated by

$$\underbrace{\text{MI}(Z,Y) - \lambda\text{MI}(Z,X)}_{\text{Expression-efficient term}} \underbrace{-\beta\text{MI}(Y,D|Z)}_{\text{ID-invariant term}}, \tag{5}$$

where $\lambda, \beta \in \mathbb{R}_+$ indicate balancing factors. Eq. 5 consists of an expression-efficient term in which Z is informative to Y and contains minimal information from X [30] and an ID-invariant term independent of D. However, the variables of Eq. 5 are intractable because they have different semantic properties and are composed of high-dimensional vectors. Each term is transformed into an operable form as follows.

1) IDENTITY-INVARIANT TERM

is expanded by the conditional and joint entropy rules: $\text{MI}(Y,D|Z) = \text{H}(Y|Z) - \text{H}(Y|D,Z)$. Here, entropy H is modeled by [45]

$$\begin{aligned} \text{H}(Y|Z) &= -\sup_g \mathbb{E}_{\boldsymbol{y},\boldsymbol{z}}\left[\log g(\boldsymbol{y}|\boldsymbol{z})\right] \\ \text{H}(Y|D,Z) &= -\sup_h \mathbb{E}_{\boldsymbol{y},\boldsymbol{z},D}\left[\log h(\boldsymbol{y}|\boldsymbol{z},D)\right]. \end{aligned} \tag{6}$$

Thanks to the universal approximation ability of neural networks [46], g and h are rearranged as target losses including parametric models:

$$\begin{aligned} \mathcal{L}_{\text{ii}} = &\ \mathbb{E}_{\boldsymbol{y}^{\text{va}},\boldsymbol{z}^{\text{exp}}}\left[\text{MSE}\left(\boldsymbol{y}^{\text{va}}, f^{\text{va}}(\boldsymbol{z}^{\text{exp}})\right)\right] \\ &- \mathbb{E}_{y^{\text{id}},\boldsymbol{z}^{\text{id}}}\left[\text{CE}\left(y^{\text{id}}, f^{\text{id}}(\boldsymbol{z}^{\text{id}})\right)\right], \end{aligned} \tag{7}$$

where VA heads f^{va} and ID head f^{id} are composed of a single FC layer, and have output dimensions of 2 and 1, respectively. MSE and CE stand for mean-squared error and cross-entropy, respectively. $\boldsymbol{z}^{\text{exp}}$ and $\boldsymbol{z}^{\text{id}}$ use $\boldsymbol{y}^{\text{va}}(\in \mathbb{R}^2)$ and $y^{\text{id}}(\in \mathbb{R}^1)$ as supervision, respectively, and provide the model with opportunities for expression regression and ID classification the same time. As a result, Eq. 7 induces to learn useful representations for $\boldsymbol{y}^{\text{va}}$ while $\boldsymbol{z}$ is robust to ID variation.

2) EXPRESSION-EFFICIENT TERM

is rewritten by Lemma 2 based on variational approximation [47] and variational information bottleneck [33]. This lemma provides a basis for converting complementary MI terms into tractable supervision loss with regularization.

Lemma 2: The term $\text{MI}(Z,Y) - \lambda\text{MI}(Z,X)$ to learn Z that contains the least information of X while being informative to Y is derived through the entropy minimized loss and the KL divergence D_{KL} with Gaussian $\mathcal{N}$ with noise ϵ.

$$\begin{aligned} \mathcal{L}_{\text{ee}} = &\ \mathbb{E}_{\boldsymbol{y}^{\text{va}},\boldsymbol{x}}\left[\text{MSE}\left(\boldsymbol{y}^{\text{va}}, f^{\text{va}}(\phi(\boldsymbol{x},\epsilon))\right)\right] \\ &+ \lambda\mathbb{E}_{\boldsymbol{x}}\left[D_{\text{KL}}[\phi(\boldsymbol{x},\epsilon)\,||\,\mathcal{N}(0,I)]\right], \end{aligned} \tag{8}$$

Proof: In general, computation of mutual information (MI) from high-dimensional vectors is intractable. Fortunately, thanks to the well-established variational information bottleneck (VIB) theory, we can show the MI-based terms $\text{MI}(Z,Y) - \lambda\text{MI}(Z,X)$ has the following lower bound (from Eq. 16 in [33]).

$$\begin{aligned} &\text{MI}(Z,Y) - \lambda\text{MI}(Z,X) \\ &\quad \geq \mathbb{E}_{\boldsymbol{y},\boldsymbol{z}}[\log q(\boldsymbol{y}|\boldsymbol{z})] - \lambda\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}\left[\log\frac{p(\boldsymbol{z}|\boldsymbol{x})}{r(\boldsymbol{z})}\right], \end{aligned} \tag{9}$$

where $p(\boldsymbol{z}|\boldsymbol{x})$ is a parametric encoder and $r(\boldsymbol{z})$ is the (variational) approximation of the true marginal. $q(\boldsymbol{y}|\boldsymbol{z})$ is considered as supervision loss to predict label $\boldsymbol{y}$ from feature $\boldsymbol{z}$. And, it is designed with cross entropy or mean-squared error. However, optimizing the second term of Eq. 9 is

[1] Although the input vectors contain different semantic factors, the FC layer will output $\boldsymbol{z}^{\text{att}}$ that integrates the properties of the vectors because of implicit neural representation [44].

still challenging. Therefore, we model $p(z|\boldsymbol{x})$ as Gaussian distribution $\mathcal{N}$ as follows.

$$p(z|\boldsymbol{x};\phi) = \mathcal{N}(z|\phi^{\mu}(\boldsymbol{x}), \phi^{\Sigma}(\boldsymbol{x})), \tag{10}$$

By attaching a single FC layer to the backbone ϕ to output the mean vector μ and the covariance matrix Σ of the same size as $\boldsymbol{z}$, respectively, ϕ^{μ} and ϕ^{Σ} are designed respectively. Then, from the reparameterization trick, this encoder is reconstructed as follows: $p(z|\boldsymbol{x})dz = p(\epsilon)d\epsilon$, where $\boldsymbol{z} = \phi(\boldsymbol{x}, \epsilon)$ is the deterministic function of $\boldsymbol{x}$ and $\epsilon \sim \mathcal{N}(0, I)$. Since this formulation offers an advantage that the noise term is independent of the model parameters, the gradient is obtained relatively easily [33]. In the end, Eq. 9 is based on the mean-squared error (MSE)-based supervision loss and the regularization loss through the KL divergence D_{KL}.

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y},\boldsymbol{z}}[\log q(\boldsymbol{y}|\boldsymbol{z})] - \lambda \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}\left[\log \frac{p(\boldsymbol{z}|\boldsymbol{x})}{r(\boldsymbol{z})}\right] \\ = \mathbb{E}_{\boldsymbol{y}^{\mathrm{va}},\boldsymbol{x}}\left[\mathrm{MSE}\left(\boldsymbol{y}^{\mathrm{va}}, f^{\mathrm{va}}(\phi(\boldsymbol{x},\epsilon))\right)\right] \\ + \lambda \mathbb{E}_{\boldsymbol{x}}\left[D_{\mathrm{KL}}[\phi(\boldsymbol{x},\epsilon) \,||\, \mathcal{N}(0, I)]\right].\end{aligned} \tag{11}$$

□

Note that in the inference phase, noise ϵ is not used: $\boldsymbol{z} = \phi(\boldsymbol{x})$.

3) FURTHER REGULARIZATION

The dependency of z^{exp} and z^{id} must be minimized for Eqs. 7 and 8 to take effect. Based on the independence testing statistics [48], we minimize the dependency between $z^{\mathrm{exp}}(\in Z^{\mathrm{exp}})$ and $z^{\mathrm{id}}(\in Z^{\mathrm{id}})$ through the following partial cross-covariance matrix.

$$\Sigma_{Z^{\mathrm{exp}},Z^{\mathrm{id}}} = \frac{1}{N-1}\sum_{i=1}^{N}\left[\left(Z_i^{\mathrm{exp}} - \frac{1}{N}\sum_{j=1}^{N} Z_j^{\mathrm{exp}}\right) \cdot \left(Z_i^{\mathrm{id}} - \frac{1}{N}\sum_{j=1}^{N} Z_j^{\mathrm{id}}\right)^{T}\right], \tag{12}$$

where N denotes the mini-batch size. While the loss in Eq. 7 pursues implicit feature learning based on model outputs (i.e., predictions), Eq. 12 is explicitly designed using features. So Eq. 12 can boost the ID-invariant effect of Eq. 7. Inspired by previous works [48], [49], $\mathcal{L}_{\mathrm{cov}}$ for disentanglement between two feature spaces is defined so that Frobenius norm $\|\Sigma_{Z^{\mathrm{exp}},Z^{\mathrm{id}}}\|_F^2$ of this matrix is minimized.

D. TOTAL OBJECTIVE AND TRAINING PROCEDURE

The total objective consists of Eqs. 7, 8, 12, and the correlation loss $\mathcal{L}_{\mathrm{corr}}$ that is widely used for boosting correlation-based metrics (cf. Sec. V-A).

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{ee}} + \alpha\mathcal{L}_{\mathrm{cov}} + \beta\mathcal{L}_{\mathrm{ii}} + \gamma\mathcal{L}_{\mathrm{corr}}, \tag{13}$$

where α, β, and γ indicate regularization coefficients. By collecting losses of different attributes, we update parameters of all neural networks except CLIP. Note that $\mathcal{L}_{\mathrm{ee}}$ is the main loss for making ϕ learn expression-efficient representations from supervision $\boldsymbol{y}^{\mathrm{va}}$. Meanwhile, $\mathcal{L}_{\mathrm{ii}}$ is a regularization loss that gives ϕ an opportunity to learn the ID-invariant representations with the demographic set $\mathcal{D}$ as a self-supervision. Since $\mathcal{L}_{\mathrm{cov}}$ disentangles the spaces of Z^{id} and Z^{exp}, it boosts the effect of $\mathcal{L}_{\mathrm{ii}}$. For details of gradient flows, see Algorithm 1.

```
Algorithm 1 Gradient Flows of EIF Framework
Require: VA head $f^{\mathrm{va}}$, ID head $f^{\mathrm{id}}$, residual module $\mathcal{R}$,
    aggregator $\mathcal{A}$, backbone $\phi$ (see Fig. 3).
Ensure: Calculate gradients of $\mathcal{L}_{\mathrm{ee}}$, $\mathcal{L}_{\mathrm{cov}}$, $\mathcal{L}_{\mathrm{ii}}$, and $\mathcal{L}_{\mathrm{corr}}$.
    $\nabla\mathcal{L}_{\mathrm{ee}}$ & $\nabla\mathcal{L}_{\mathrm{corr}}$ {min $f^{\mathrm{va}}$ and $\phi$.}
    $\nabla\mathcal{L}_{\mathrm{cov}}$ {min $\mathcal{R}$, $\mathcal{A}$, and $\phi$.}
    $\nabla\mathcal{L}_{\mathrm{ii}}$ {max $f^{\mathrm{id}}$ & min $f^{\mathrm{va}}$, $\mathcal{R}$, $\mathcal{A}$, $\phi$.}
    =0
```

V. EXPERIMENTS

A. SETTINGS

1) CONFIGURATIONS

The proposed model was implemented with PyTorch library [50], and it was trained at Intel Xeon CPU and NVIDIA RTX A100 GPU. Every quantitative value is the average of the results of five experiments. The parameters of convolutional and FC layers were updated through Adam optimizer [51] with learning rate (LR) 4×10^5. LR decreases by 0.8 times in the initial 5K iteration, and then decreases by 0.8 times every 20K iteration. Three backbones (ϕ) were selected: AlexNet(AL) [18], ResNet18(R18) [19], and Mlp-Mixer(MMx) [52]. Here, the parameter-reduced AlexNet-tuned was employed [8], [12]. Mini-batch sizes (N) of the backbones were set to 512, 256, and 128, respectively. τ of Eq. 3 was 0.5 and λ of Eq. 8 was 10^3. α, β, and γ of Eq. 13 were set to 10^3, 10^5, and 0.5, respectively. The dimensions of $\boldsymbol{z}$, $\boldsymbol{v}$, and $\boldsymbol{t}^{\mathcal{D}}$ were all 512.

2) MODEL DETAILS

EIF framework consists of 5 modules: backbone ϕ, aggregator $\mathcal{A}$, residual module $\mathcal{R}$, VA head f^{va}, and ID head f^{id}. First, as ϕ, the default version ResNet18 [19], Mlp-Mixer [52], and parameter reduced AlexNet [18] were adopted. In detail, AlexNet's convolution layer is the same as the original, but utilizes two lightweight FC layers with ReLU activation as follows: $\mathsf{FC}(32) - \mathsf{FC}(512)$, where the parenthesis indicates the size of the output vector. $\mathcal{A}$ and $\mathcal{R}$ are designed as $\mathsf{FC}(512)$. Finally, f^{va} and f^{id} are designed as $\mathsf{FC}(2)$ and $\mathsf{FC}(1)$, respectively.

3) BASELINE

Two open source techniques, i.e., CAF [12] and ELIM [8], were used as key baselines in the experiments.

4) PUBLIC DATASETS

Aff-wild [6] is a video database containing the reactions of subjects in TV shows and soap operas. About 1.2M frames extracted from a total of 298 videos are divided into train-set

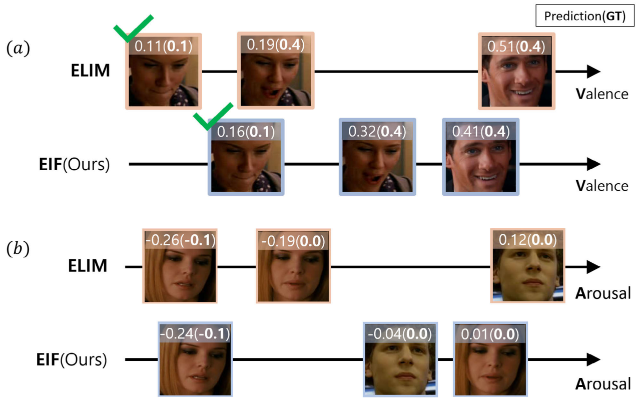


FIGURE 6. Example-based results demonstrating ID-invariant representation capabilities. (*a*) and (*b*) denote Valence and Arousal, respectively.

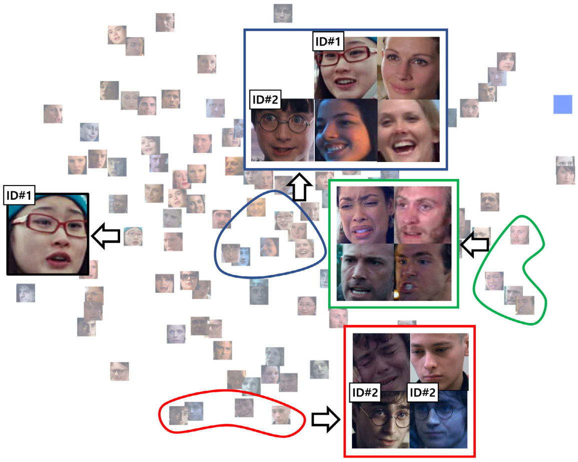


FIGURE 7. Visualization of expression features using t-SNE [9].

and validation set [8], [12]. Aff-wild2 [53] is an extended version of Aff-wild in which spontaneous facial expressions of subjects and background variations (e.g., illumination) are additionally considered. About 1.4M frames were extracted from the videos and divided into train- and validation-sets. AFEW-VA [5] is a movie database composed of 600 short video clips, and the cross-validation scheme of 5:1 ratio is adopted. Note ID is assigned from the $\{\text{folder}_{\#}\}$ per video as in [8].

5) EVALUATION METRICS

Given the emotion label set Y and the prediction set $\hat{Y}$, the four metrics used in model inference of EIF and other techniques are as follows.

▶ Root mean-squared error (RMSE) measures the average of the squares of the errors, i.e., the straight difference between predictions and labels.

$$\mathrm{RMSE}(Y, \hat{Y}) = \sqrt{\mathbb{E}((Y - \hat{Y})^2)}.$$

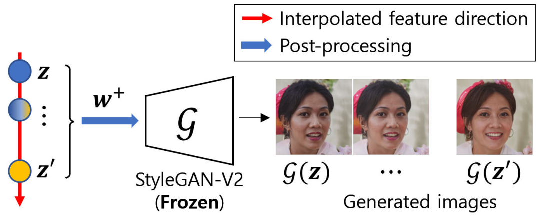


FIGURE 8. Pipeline of feature-extrapolated image generation task.

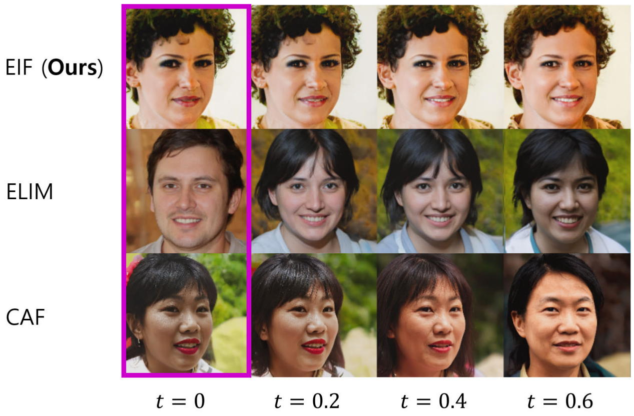


FIGURE 9. Generated traversal images. The magenta box contains the original (i.e., not extrapolated) images.

▶ Pearson correlation coefficient (PCC) measures the linear correlation between predictions and labels. In words, the PCC summarizes the characteristics of two different (data)sets.

$$\mathrm{PCC}(Y, \hat{Y}) = \frac{\mathbb{E}[(Y - \mu_Y)(\hat{Y} - \mu_{\hat{Y}})]}{\sigma_Y \sigma_{\hat{Y}}},$$

where μ and σ indicate the mean and standard deviation of the test samples, respectively.

▶ Concordance correlation coefficient (CCC) measures agreement between two sets. That is, the CCC is the addition of the bias correlation factor to the PCC.

$$\mathrm{CCC}(Y, \hat{Y}) = \frac{2\sigma_Y \sigma_{\hat{Y}} \mathrm{PCC}(Y, \hat{Y})}{\sigma_Y^2 + \sigma_{\hat{Y}}^2 + (\mu_Y - \mu_{\hat{Y}})^2}.$$

▶ Sign agreement (SAGR) determines whether the signs of predictions and labels match based on the VA axes.

$$\mathrm{SAGR}(Y, \hat{Y}) = \frac{1}{N^{te}} \sum_{i=1}^{N^{te}} \Gamma(\mathrm{sign}(\boldsymbol{y}_i^{\mathrm{va}}), \mathrm{sign}(\hat{\boldsymbol{y}}_i^{\mathrm{va}})),$$

where $\boldsymbol{y}^{\mathrm{va}} \in Y$ and $\hat{\boldsymbol{y}}^{\mathrm{va}} \in \hat{Y}$. N^{te} stands for the number of images in the test set. Γ is a binary function that outputs 1 if two values have the same sign, and 0 otherwise.

In addition, the correlation loss $\mathcal{L}_{\mathrm{corr}}$ for model training is defined as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{corr}} &= \mathcal{L}_{\mathrm{PCC}}(Y, \hat{Y}) + \mathcal{L}_{\mathrm{CCC}}(Y, \hat{Y}) \\ &= \left(1 - \frac{\mathrm{PCC}_v(Y, \hat{Y}) + \mathrm{PCC}_a(Y, \hat{Y})}{2}\right) \\ &\quad + \left(1 - \frac{\mathrm{CCC}_v(Y, \hat{Y}) + \mathrm{CCC}_a(Y, \hat{Y})}{2}\right), \qquad (14)\end{aligned}$$

TABLE 1. Realism ratings of 18 trained volunteers. Columns 1-4 show the number of times that users gave this rating. The column "real" shows the percentage of users that rated the images with 3 or 4.

	EIF (**Ours**)					ELIM [8]					CAF [12]				
	1	2	3	4	'real'	1	2	3	4	'real'	1	2	3	4	'real'
Images in Fig. 9	2	1	2	13	83%	12	5	1	-	6%	9	9	-	-	0%
#1 (Children)	4	-	12	2	78%	10	5	3	-	17%	9	9	-	-	0%
#2 (Men)	-	-	3	15	100%	1	6	9	2	61%	-	-	4	14	100%
#3 (Women)	-	-	3	15	100%	3	9	5	1	33%	13	5	-	-	0%

where PCC(/CCC)$_v$ and PCC(/CCC)$_a$ indicate correlation values on the valence and arousal axes, respectively.

6) SETTING FOR USER STUDY

The goal of user study is to observe the changes in facial expressions while maintaining ID attributes. Through this experiment, we can see how well the features learn the ID-invariant representation ability in high-dimensional image space. Since separate attribute classifiers are not given, the realism rating of the generated image was subjectively evaluated.

B. IDEA VERIFICATION

In order to clearly evaluate the validity of the proposed method, this section is described in the form of answering the following key questions. Q1. Is EIF valid for ID dependency cases? Q2. Do features show ID-invariant ability?

A1. Effectiveness for ID dependency. The goal of I^{2}FER is to make it robust against the so-called ID shift phenomenon [8] where face-related attributes are shifted while following the same label prediction mechanism. We compared the FER performance of ELIM [8] targeting I^{2}FER and EIF through several ID dependency cases. First, to demonstrate ID-invariant representation ability in terms of valence, we compared the emotional tendencies of two randomly selected ID samples. ELIM estimated GT values more elaborately than EIF (see green symbols in Fig. 6(a)), but showed results that were somewhat more dependent on ID than similar expressions. On the other hand, EIF was independent of ID and showed fairly sophisticated estimation results. The same experimental results were obtained in the arousal axis. As a result, EIF showed prediction results that were less dependent on ID changes in the negative region of the arousal axis, where estimation difficulty was highest (see Fig. 6(b)).

A2. Verifying ID-invariant ability. The representation ability of EIF is evaluated by visualizing z^{exp} isolated from $\mathcal{R}$ (of Sec. IV-B). Figure 7 illustrates the correlation between different facial expressions in the feature level. We can find that the samples are not clustered depending on IDs but are distributed depending on the expression changes. For example, the sample of ID#2 with a smiling expression is located near the ID#1 sample with a similar expression (see the blue box), and the samples of ID#2 with a sad expression are gathered on the bottom (see the red box). This demonstrates the validity of $\mathcal{R}$ separating z^{exp} from z and shows the regularization effect of $\mathcal{L}_{\text{ii}}$.

TABLE 2. Results on the validation set of Aff-wild2. The gray background represents the results of the proposed method, i.e., EIF. † stands for in-house implementation.

Methods	RMSE-V (↓)	RMSE-A (↓)	CCC-V (↑)	CCC-A (↑)
Self-Attention	0.374†	0.273†	0.419	0.505
AP	-	-	0.438	0.498
ELIM (AL)	0.357†	0.256†	0.451	0.478
EIF (AL)	0.324	0.232	0.504	0.500
ELIM (R18)	0.349†	0.243†	0.449	0.496
EIF (R18)	0.320	0.230	0.427	0.534
ELIM (MMx)	0.332†	0.210†	0.498	0.493
EIF (MMx)	**0.297**	**0.203**	**0.548**	**0.531**

C. DOWNSTREAM TASK

1) PRACTICAL ASSUMPTION

We were inspired by a well-known study [54] that demonstrated applications such as (facial) expression transfer by re-embedding user-specified images into the StyleGAN latent space. That is, StyleGAN(-V2)'s latent space [55] does not need to be set to a *random vector*. After converting z to extended latent space $\boldsymbol{w}^+ \in \mathbb{R}^{18\times 512}$, we synthesize traversal images through feature extrapolation.

z of the model trained with emotional labels assumes that only the expression attributes change in the manifold. So, we pay attention to the feature-extrapolated image generation task where the attributes implied by z can be compared in high-dimensional image space ($\in \mathbb{R}^{1024\times 1024}$). Through this task, the above assumption and ID-invariant representation are further verified. The pipeline is shown in Fig. 8. First, a set of individuals S is built by performing feature extrapolation toward a specific direction.

$$S(z) = \left\{z' := z + t \cdot \boldsymbol{a} \mid t \in [-1, 1]\right\},$$

where t stands for a scale factor. Here, the attribute vector $\boldsymbol{a}$ that determines the extrapolation direction is designed as a simple one vector $\mathbf{1}$. This assumes that all elements of $\boldsymbol{a}$ change almost equally. Then, the pre-trained generator $\mathcal{G}$ (e.g., StyleGAN-V2 [55]) generates images with the extrapolated features $z'(\in S)$ and z as two inputs. If only the expression changes smoothly while maintaining the ID, it is considered as an ID-invariant case. Otherwise, it is regarded as an ID-dependent case. This decision criterion is formulated as follows:

$$\text{ID-shifting}(\mathcal{G}(z), \mathcal{G}(z')) = 0 \rightarrow \text{ID invariant},$$
$$\text{ID-shifting}(\mathcal{G}(z), \mathcal{G}(z')) = 1 \rightarrow \text{ID dependent}.$$

Figure 9 qualitatively compares EIF and prior arts. CAF and ELIM showed a change in ID rather than a change in expression due to abrupt changes in gender and skin color (ID-shifting = 1). On the other hand, in EIF, only the expression change was mainly observed while the ID attribute was preserved (ID-shifting = 0).

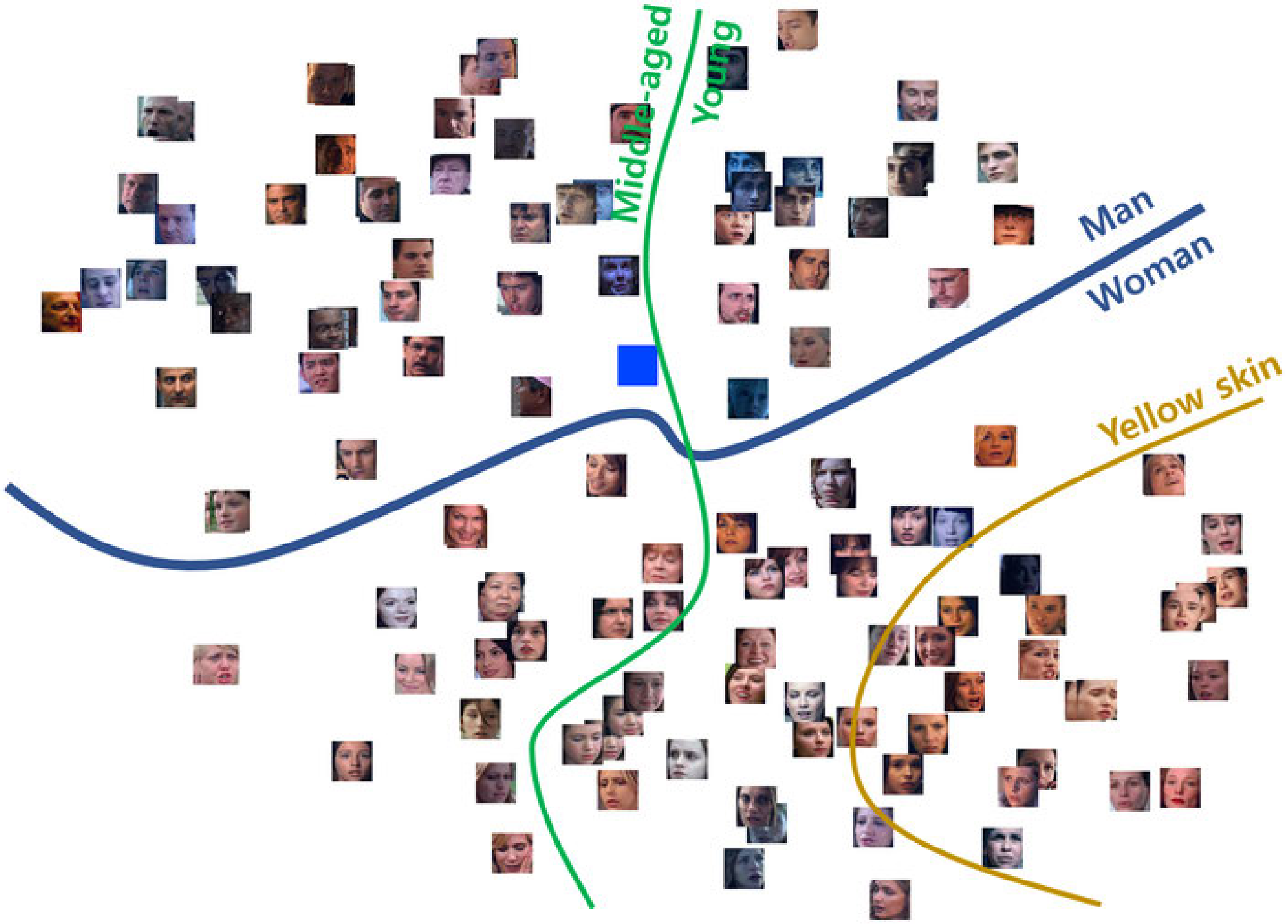


FIGURE 10. t-SNE visualization [9] of ID feature z^{id}. We observed a tendency for samples to cluster in a low-dimensional space based on the three factors that make up the ID attribute.

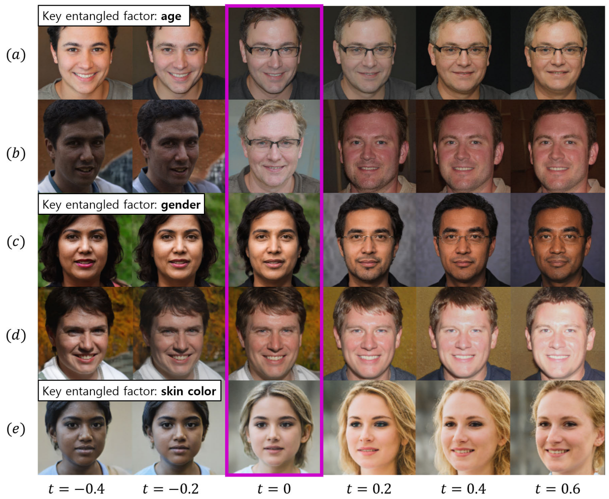

Key entangled factor: age
Key entangled factor: gender
Key entangled factor: skin color
(a)
(b)
(c)
(d)
(e)
$t = -0.4$
$t = -0.2$
$t = 0$
$t = 0.2$
$t = 0.4$
$t = 0.6$

FIGURE 11. Generated traversal images from ID feature z^{id}. The magenta box contains the original (i.e., not extrapolated) images. Due to the file size limitation, the resolution of the generated image is partially reduced. $t (\in [-1, 1])$ stands for scale factor.

Next, the generation results were verified through user study that rates the realism of each image. This experiment was designed because ID classifiers for ID-shifting [56], [57] are not normally available. In Table 1, EIF improved by 77% and 83% in realism score than ELIM and CAF, respectively. Refer to the next two paragraphs for the other generation results with z^{id} and in-depth analysis of ID attributes.

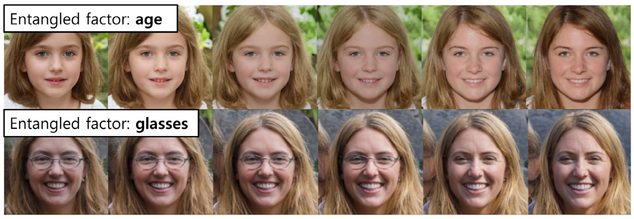


FIGURE 12. Failure cases that contain factors other than expression.

2) VISUALIZATION OF ID FEATURE

We verify the training validity of z^{id} by visualizing ID features (z^{id}) in a low-dimensional space. Fig. 10 visualizes z^{id} through t-SNE. First, samples tend to be located depending on two factors of gender (attribute), i.e., man and woman. The relative positions of the samples also seem to be determined

TABLE 3. Experimental results on AFEW-VA. Unlike 'Temporal', 'Static' trains the model based only on a single image. According to the regime of parameter size, 'Heavy', 'Middle', and 'Light' are divided: over 20M, more than 10M and less than 20M, and less than 5M.

Case	Type	Methods	RMSE (↓)		SAGR (↑)		PCC (↑)		CCC (↑)	
			(V)	(A)	(V)	(A)	(V)	(A)	(V)	(A)
Static	Light	Kossaifi *et al.*	0.24	0.24	0.64	0.77	0.55	0.57	0.55	0.52
		CAF (AL)	0.20	0.20	0.66	0.83	0.67	0.63	0.58	0.57
		ELIM (AL)	0.186	0.198	0.692	0.815	0.680	0.645	0.602	0.581
		EIF (AL)	0.161	0.167	0.776	0.784	**0.731**	0.726	**0.689**	**0.694**
	Middle	Mitenkova *et al.*	0.40	0.41	-	-	0.33	0.42	0.33	0.40
		CAF (R18)	0.17	0.18	0.68	0.87	0.67	0.60	0.59	0.54
		ELIM (R18)	0.168	0.170	0.723	0.876	0.651	0.679	0.615	0.614
		EIF (R18)	0.155	0.152	0.793	0.877	0.623	**0.775**	0.621	0.712
	Heavy	ELIM (MMx)	0.167	0.164	0.747	0.877	0.704	0.707	0.643	0.598
		EIF (MMx)	**0.151**	**0.142**	**0.812**	**0.893**	0.715	0.764	0.656	0.669
Temporal	Light	HO-Conv	0.28	0.19	0.53	0.75	0.12	0.23	0.11	0.15
		HO-Conv-trans.	0.20	0.21	0.67	0.79	0.64	0.62	0.57	0.56
	Heavy	Kollias *et al.*	-	-	-	-	0.51	0.58	0.52	0.56

TABLE 4. Performance comparison on the Aff-wild dataset.

Type	Methods	Backbone	PCC (↑)		CCC (↑)	
			(V)	(A)	(V)	(A)
Light	Hasani et al. [20]	ResNeXt50	0.42	0.40	0.37	0.31
	CAF [12]	AlexNet	0.55	0.57	0.54	0.56
	ELIM [8]		0.63	0.64	0.63	0.61
	EIF		**0.68**	0.61	**0.64**	0.59
Middle	CAF [12]	ResNet18	0.57	0.57	0.55	0.56
	ELIM [8]		0.64	0.60	0.61	0.57
	EIF		0.65	0.66	0.61	0.60
Heavy	Hasani et al. [58]	Inception-Res.	0.44	0.26	0.36	0.19
	ELIM [8]	Mlp-Mixer	0.65	0.68	0.61	0.63
	EIF		0.67	**0.69**	0.63	**0.64**

TABLE 5. Ablation study on Aff-wild2. Aside from the attributes and losses, the other configurations remained the same.

CRL formula	Appearance	Gender	Skin color	Age	CCC (↑)	
					(V)	(A)
EIF (MMx)	✓				0.488	0.466
	✓	✓			0.500	0.512
	✓		✓		0.491	0.503
	✓	✓	✓		0.525	0.528
EIF (MMx) (Tab. 2)	✓	✓	✓	✓	0.548	0.531
w/o $\mathcal{L}_{\text{cov}}$	✓	✓	✓	✓	0.529	0.517
w/o $\mathcal{L}_{\text{corr}}$	✓	✓	✓	✓	0.512	0.491
w/o $\mathcal{L}_{\text{cov}}$ & $\mathcal{L}_{\text{corr}}$	✓	✓	✓	✓	0.451	0.457

by the factors of age (attribute). However, the influence was not greater than that of gender. In fact, this result is closely related to CLIP's encoding ability. Since gender is a superficial attribute on faces, its feature encoding is relatively easy. However, since age is difficult to predict directly from faces and is an attribute with different tendencies for each person, its encoding is not easy.

3) TRAVERSAL IMAGES FROM ID FEATURE

Finally, traversal images were generated by inputting ID features and extrapolated ID features to the pre-trained generator $\mathcal{G}$. The goal of this downstream task is to observe whether the transitions of gender, skin, and age attributes are also represented in high-dimensional image space. Note that this experiment indirectly confirms the factorization performance of the residual module $\mathcal{R}$, which explicitly separates expression-dependent and ID-dependent components. Examples of traversal images generated by the pipeline in Fig. 6 of the main body are shown in Fig. 11 here. The results of the third column are images generated from z^{id} (see the magenta box). And, the images generated from the ID features extrapolated toward the positive or negative direction are shown row-by-row in the figure. Age (see Fig. 11(*a*)), gender (see Fig. 11(*c*)), and skin color (see Fig. 11(*e*)) were the main components that changed the appearance of traversal images. Surprisingly, variations in other attributes (e.g., glasses, head pose, beard etc.) that we did not encode were also observed as in Figs. 11(*a*) and (*b*)). These examples show that z has many face-related attributes that we didn't consider. This generative result, which shows a much more natural age progression or gender transition than the latest studies, demonstrates the scalability of EIF, that is, EIF can be extended in many ways to face-related generation tasks. In the future, if more sophisticated attribute encoding or feature manipulation methods are combined with this downstream task, more natural facial expression changes can be produced.

In the Aff-wild, which is often used to verify the affect estimation ability, EIF successfully captured the emotional change trend. In Table 4, EIF achieved PCC-V, about 5% better than ELIM in the light model. Also, thanks to the high capacity of Mlp-Mixer, EIF achieved SOTA performance in the Arousal axis.

D. COMPARISON WITH PRIOR ARTS

This section evaluates the superiority of EIF in quantitative aspects. Table 2 compares the proposed method with prior arts for the large-scale Aff-wild2. EIF achieved RMSE and CCC significantly ahead of stochastic process-based AP [7] and optimal ID matching-based ELIM [8]. For example, EIF (AL) showed about 0.03 improvement in RMSE-V than ELIM (AL), and EIF (MMx) showed 5% improvement in CCC-V than ELIM (MMx). Considering that Aff-wild2 contains even sudden emotional changes, this performance improvement is quite significant.

Table 3 shows the spontaneous expression estimation performance in terms of learning method, model size, and four evaluation metrics. ELIM [8] and EIF focusing on I^{2}FER lead the way. This ranking suggests that the ID-invariant mechanism is more effective for affect estimation in VA space than temporal information. Overall, the performance improved as the model size increased. However, EIF (AL) showed a 3.3% improvement in CCC-V than EIF (MMx). This demonstrates the strength of EIF, which can perform FER well even with a light size model.

In the Aff-wild, which is often used to verify the affect estimation ability, EIF successfully captured the emotional change trend. In Table 4, EIF achieved PCC-V, about 5% better than ELIM in the light model. Also, thanks to the high capacity of Mlp-Mixer, EIF achieved SOTA performance in the Arousal axis.

E. ABLATION STUDY

We answer the following questions regarding the composition of EIF. Q3. Does the number of inferred attributes affect performance? Q4. Are the losses for regularization effective? Q5. Are attributes other than facial expressions encoded?

A3. Impact of attributes. The more face-related attributes are observed, the more accurately the subject ID can be recognized [59]. In order to observe that this claim works in EIF, we investigated the change in performance for each element of $\mathcal{D}$ (see Table 5). When gender was added based on $\boldsymbol{v}$ from $\mathcal{E}_v$ (i.e., appearance), a large performance improvement of 4.6% was observed in CCC-A. On the other hand, the addition of age showed only 0.3% increase in CCC-A. This indirectly verifies CLIP's weak age encoding ability.

A4. Impact of losses. Due to the nature of EIF, disentanglement between two feature spaces is very important. So, we analyze the influence of $\mathcal{L}_{\text{cov}}$ (cf. Eq. 12) performing the disentanglement learning from a cross-covariance matrix. As in the 6th row of Table 5, the absence of $\mathcal{L}_{\text{cov}}$ showed a 1.9% CCC-V decrease. In the absence of $\mathcal{L}_{\text{corr}}$ boosting the learning of emotional tendency, a decrease in CCC-A of 4% was observed. Also, when all of the regularization losses are removed, a large performance decrease of about 10% in CCC-V was observed.

A5. Incomplete disentanglement. The powerful encoding capability of CLIP helps EIF successfully extract the attributes of subjects. However, since z^{att} is obtained from only three factors (i.e., gender, age, skin color), redundant attribute(s) that are not separated from z can exist. As in the traversal images of Fig. 12, z^{exp} includes some attributes (i.e., glasses and age) that are not related to expressions. Age attribute is still observed. Note that [60] has officially stated that CLIP is relatively weak in age estimation [60]. The first row of Fig. 12 contains the subject's smooth age progress as well as the expression change. Our result showing a much more natural subject growth than recent face aging works [61], [62] demonstrates the significant potential of EIF, which can expand into face-related generation tasks.

TABLE 6. Comparison with 7-classes FER methods.

Methods	Backbone	Accuracy (%)	
		AffectNet	RAF-DB
DLN [40]	Inception-Res.	63.70	86.40
RAN [63]	ResNet-18	59.50	86.90
SCN [64]		60.23	87.03
IPD-FER [65]		62.23	88.89
DMUE [66]		62.84	88.76
EIF		64.37	88.93
DMUE [66]	ResNet-50	63.11	89.42
Face2Exp [67]		64.23	88.54
EIF		**64.85**	**89.44**

A6. 7-classes FER results. To validate the superior performance of our method, we compare its performance in a 7-classes FER (discrete FER) configuration using the AffectNet [17] and RAF-DB [68]. For the experiment, we substitute $\mathcal{L}_{\text{ee}}$ with cross-entropy and remove $\mathcal{L}_{\text{ii}}$ due to the nature of AffectNet and RAF-DB, where multiple images for a single identity are not available. The experimental results are given in Table 6. When compared to the state-of-the-art discrete facial FER technique, Face2Exp [67], the proposed method demonstrates performance improvements of 0.62% on AffectNet and 0.90% on RAF-DB. In addition, when compared to IPD-FER, which employs the same backbone, the proposed method shows a performance improvement of 2.14% on AffectNet. Note that among the existing methods, DLN [40] and IPD-FER [65] explicitly targeted I^{2}FER. Furthermore, even when using a relatively lightweight backbone, the proposed method outperforms DLN by 0.67% on RAF-DB. Therefore, experimentally, the proposed method targeting I^{2}FER has been demonstrated to achieve superior performance in both continuous (i.e., VA) FER and discrete FER.

VI. CONCLUSION

Towards ID-invariant representation, we propose a novel FER framework that uses ID and expression features as inputs for complementary MI-based terms. ID-invariant ability was experimentally verified in both low-dimensional space (through t-SNE) and high-dimensional image space. This paper will give considerable insight to future FER studies in that it successfully deals with generalization against ID dependency, which is the key to VA FER.

Potential Societal Impact: The proposed method to infer the attributes of images with CLIP may cause privacy or surveillance-related threats. For example, searching for attributes of celebrities and analyzing their facial expressions can be a social threat in terms of privacy. So it is urgent to introduce privacy-aware attribute searching or differential privacy.

REFERENCES

[1] M. Khan, W. Gueaieb, A. El Saddik, and S. Kwon, "MSER: Multimodal speech emotion recognition using cross-attention with deep fusion," *Expert Syst. Appl.*, vol. 245, Jul. 2024, Art. no. 122946.

[2] B. C. Song and D. H. Kim, "Hidden emotion detection using multi-modal signals," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–7.

[3] P. Ekman, "Basic emotions," *Handbook Cognition Emotion*, vol. 98, nos. 45–60, p. 16, 1999.

[4] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Personality*, vol. 11, no. 3, pp. 273–294, Sep. 1977.

[5] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "AFEW-VA database for valence and arousal estimation in-the-wild," *Image Vis. Comput.*, vol. 65, pp. 23–36, Sep. 2017.

[6] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-wild: Valence and arousal 'in-the-wild' challenge," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1980–1987.

[7] E. Sanchez, M. K. Tellamekala, M. Valstar, and G. Tzimiropoulos, "Affective processes: Stochastic modelling of temporal context for emotion and facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9070–9080.

[8] D. Kim and B. C. Song, "Optimal transport-based identity matching for identity-invariant facial expression recognition," in *Proc. NeurIPS*, 2022, pp. 18749–18762.

[9] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

[10] V. Bruce and A. Young, "Understanding face recognition," *British J. Psychol.*, vol. 77, no. 3, pp. 305–327, 1986.

[11] C. Darwin and P. Prodger, *The Expression of the Emotions in Man and Animals*. London, U.K.: Oxford Univ. Press, 1998.

[12] D. H. Kim and B. C. Song, "Contrastive adversarial learning for person independent facial emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 5948–5956.

[13] K. Ali and C. E. Hughes, "Facial expression recognition by using a disentangled identity-invariant expression representation," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 9460–9467.

[14] C. Jia, M. Luo, Z. Dang, X. Chang, and Q. Zheng, "Towards real-time person search with invariant feature learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[15] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via IBN-Net," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 464–479.

[16] A. Radford, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[17] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[20] B. Hasani, P. S. Negi, and M. H. Mahoor, "BReG-NeXt: Facial affect computing using adaptive residual networks with bounded gradient," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 1023–1036, Apr. 2022.

[21] P. Barros, G. Parisi, and S. Wermter, "A personalized affective memory model for improving emotion recognition," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 485–494.

[22] Y. Zhao, X. Shen, Z. Jin, H. Lu, and X.-s. Hua, "Attribute-driven feature disentangling and temporal aggregation for video person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4908–4917.

[23] G. Wang, H. Han, S. Shan, and X. Chen, "Cross-domain face presentation attack detection via multi-domain disentangled representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6677–6686.

[24] Z. Chen, Y. Luo, R. Qiu, S. Wang, Z. Huang, J. Li, and Z. Zhang, "Semantics disentangling for generalized zero-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8692–8700.

[25] T. H. Nguyen-Phuoc, C. Richardt, L. Mai, Y. Yang, and N. Mitra, "BlockGAN: Learning 3D object-aware scene representations from unlabelled images," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6767–6778.

[26] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 6, Jun. 2004, Art. no. 066138.

[27] M. Boudiaf, I. M. Ziko, J. Rony, J. Dolz, P. Piantanida, and I. B. Ayed, "Information maximization for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst. Annu. Conf. Neural Inf. Process. Syst.*, Dec. 2020, pp. 2445–2457.

[28] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[29] P. Savarese, S. S. Y. Kim, M. Maire, G. Shakhnarovich, and D. McAllester, "Information-theoretic segmentation by inpainting error maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4028–4038.

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

[31] Z. Wan, C. Zhang, P. Zhu, and Q. Hu, "Multi-view information-bottleneck representation learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10085–10092.

[32] J. Lee, J. Choi, J. Mok, and S. Yoon, "Reducing information bottleneck for weakly supervised semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 27408–27421.

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

[34] G. Gao, H. Huang, C. Fu, Z. Li, and R. He, "Information bottleneck disentanglement for identity swapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3403–3412.

[35] Z. Dang, M. Luo, C. Jia, G. Dai, J. Wang, X. Chang, J. Wang, and Q. Zheng, "Disentangled representation learning with transmitted information bottleneck," 2023, *arXiv:2311.01686*.

[36] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised speech decomposition via triple information bottleneck," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7836–7846.

[37] W. Zhang, X. Ji, K. Chen, Y. Ding, and C. Fan, "Learning a facial expression embedding disentangled from identity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6755–6764.

[38] J. Pearl, "Causal inference in statistics: An overview," *Statist. Surv.*, vol. 3, pp. 96–146, Jan. 2009.

[39] Y. Chen, D. Chen, T. Wang, Y. Wang, and Y. Liang, "Causal intervention for subject-deconfounded facial action unit recognition," 2022, *arXiv:2204.07935*.

[40] Z. Wang, Y. Luo, R. Qiu, Z. Huang, and M. Baktashmotlagh, "Learning to diversify for single domain generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 814–823.

[41] J. Cha, K. Lee, S. Park, and S. Chun, "Domain generalization by mutual-information regularization with pre-trained models," 2022, *arXiv:2203.10789*.

[42] A. Asadi, E. Abbe, and S. Verdu, "Chaining mutual information and tightening generalization bounds," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[43] K. Kärkkäinen and J. Joo, "FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1547–1557.

[44] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Proc. NIPS*, 2020, pp. 7462–7473.

[45] F. Farnia and D. Tse, "A minimax approach to supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–17.

[46] F. Scarselli and A. C. Tsoi, "Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results," *Neural Netw.*, vol. 11, no. 1, pp. 15–37, Jan. 1998.

[47] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent.*, 2013.

[48] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, ''Kernel measures of conditional dependence,'' in *Proc. Adv. neural Inf. Process. Syst.*, vol. 20, 2007.

[49] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He, and Z. Shen, ''Deep stable learning for out-of-distribution generalization,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5372–5382.

[50] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga, ''PyTorch: An imperative style, high-performance deep learning library,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[51] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015.

[52] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, and J. Uszkoreit, ''MLP-mixer: An all-MLP architecture for vision,'' *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 24261–24272, 2021.

[53] D. Kollias and S. Zafeiriou, ''Expression, affect, action unit recognition: Aff-Wild2, multi-task learning and arcface,'' in *Proc. Brit. Mach. Vis. Conf.*, 2019, p. 297.

[54] R. Abdal, Y. Qin, and P. Wonka, ''Image2StyleGAN: How to embed images into the StyleGAN latent space?'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4431–4440.

[55] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, ''Analyzing and improving the image quality of StyleGAN,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.

[56] R. Denton, B. Hutchinson, M. Mitchell, T. Gebru, and A. Zaldivar, ''Image counterfactual sensitivity analysis for detecting unintended bias,'' 2019, *arXiv:1906.06439*.

[57] M. Peychev, A. Ruoss, M. Balunovic, M. Baader, and M. Vechev, ''Latent space smoothing for individually fair representations,'' 2021, *arXiv:2111.13650*.

[58] B. Hasani and M. H. Mahoor, ''Facial affect estimation in the wild using deep residual and convolutional networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1955–1962.

[59] A. J. O'Toole, P. J. Phillips, X. An, and J. Dunlop, ''Demographic effects on estimates of automatic face recognition performance,'' *Image Vis. Comput.*, vol. 30, no. 3, pp. 169–176, Mar. 2012.

[60] (Jan. 5, 2021). *CLIP: Connecting Text and Images*. [Online]. Available: https://openai.com/blog/clip/

[61] Z. Li, R. Jiang, and P. Aarabi, ''Continuous face aging via self-estimated residual age embedding,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15003–15012.

[62] F. Makhmudkhujaev, S. Hong, and I. K. Park, ''Re-aging GAN: Toward personalized face age transformation,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3888–3897.

[63] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, ''Region attention networks for pose and occlusion robust facial expression recognition,'' *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.

[64] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, ''Suppressing uncertainties for large-scale facial expression recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6896–6905.

[65] J. Jiang and W. Deng, ''Disentangling identity and pose for facial expression recognition,'' *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 1868–1878, Oct. 2022.

[66] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, ''Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6244–6253.

[67] D. Zeng, Z. Lin, X. Yan, Y. Liu, F. Wang, and B. Tang, ''Face2Exp: Combating data biases for facial expression recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Washington, DC, USA: Combat, Jun. 2022, pp. 20259–20268.

[68] S. Li, W. Deng, and J. Du, ''Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2584–2593.

DAEHA KIM received the B.S. degree in electronic engineering and the Ph.D. degree in electrical and computer engineering from Inha University, Incheon, South Korea, in 2017 and 2023, respectively. His research interests include facial expression recognition, expression-based talking face generation, and large language models (LLMs).

SEONGHO KIM (Associate Member, IEEE) received the B.S. degree in electronic engineering from Inha University, Incheon, South Korea, in 2022, where he is currently pursuing the M.S. degree in electrical and computer engineering. His research interests include facial expression manipulation and video synthesize.

BYUNG CHEOL SONG (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 1994, 1996, and 2001, respectively. From 2001 to 2008, he was a Senior Engineer with Samsung Research (formerly, Digital Media Research and Development Center), Samsung Electronics Company Ltd., Suwon, South Korea. In 2008, he joined the Department of Electronic Engineering, Inha University, Incheon, South Korea, where he is currently a Professor. His research interests include image processing and computer vision.