This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Decoupling Identity Confounders for Enhanced Facial Expression Recognition: An Information-Theoretic Approach

Mohd Aquib
IIT Kanpur, India
aquib@iitk.ac.in

Nishchal K. Verma
IIT Kanpur, India
nishchal@iitk.ac.in

M. Jaleel Akhtar
IIT Kanpur, India
mjakhtar@iitk.ac.in

Abstract

Facial expression recognition (FER) remains challenging due to subtle inter-class variations and significant intra-class differences, often exacerbated by identity-specific features confounding the expression features. While recent methods attempt to disentangle identity from expression, they often rely on auxiliary labels or computationally expensive image generation, limiting scalability. To address this, we propose DICE-FER (Decoupling Identity Confounders for Enhanced FER), a novel framework that decouples identity confounders from expression features through mutual information (MI) estimation without requiring labels or reconstruction. DICE-FER processes paired images with shared expressions, partitioning their features into (1) expression representations which is maximized via cross-referenced MI and (2) identity representations which is adversarially minimized for MI with expression. This dual optimization isolates identity-invariant expression cues while eliminating the need for costly generation or subject annotation. Experiments on benchmark datasets demonstrate that DICE-FER outperforms state-of-the-art methods in both disentanglement quality and recognition accuracy.

1. Introduction

Facial expressions are a cornerstone of nonverbal communication, driving advancements in facial expression recognition (FER) for applications ranging from mental health diagnostics to human-robot interaction. Despite progress, FER systems remain hindered by identity-related factors, which act as a confounding variable that amplifies two core challenges: the fine-grained distinctions between expression classes (e.g., a smile vs. a smirk) and pronounced intra-class variability across individuals (e.g., happiness expressed by different demographics). Unlike transient exogenous factors like pose or lighting, identity systematically biases learned representations by conflating static facial attributes (e.g., bone structure, demographic traits) with dynamic expression semantics. Although recent methods have aimed to tackle these issues, many still struggle to separate exogenous identity factors efficiently, complicating the extraction of pure expression features [36].

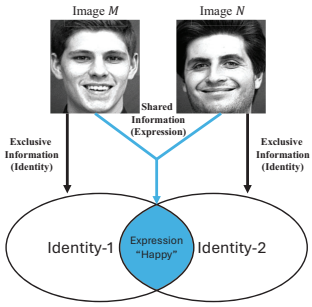


Fig. 1. Given images *M* and *N*, we aim to learn a feature space that decouples the shared information (expression) from the exclusive information (identity) of the image.

Disentangling expression features from other confounding factors has emerged as a promising strategy to amplify FER performance. Some methods [41], [42], [4], [38], [8] attempt to disentangle expression-related features from identity-related ones, complete disentanglement remains a challenge due to the overlap between these feature sets. For instance, TDGAN [39] employs a generative adversarial strategy to separate expression information, while DDL [32] leverages multitask and adversarial learning to simultaneously remove multiple interfering factors. Li *et al.* [22] use HSIC contrastive loss in a Siamese network to achieve identity invariant FER, and IPD-FER [17] decouples identity and posture to isolate expression features. DR-ALNet [10] uses a combinatorial loss function to isolates expression features from non-expression features but it may struggle to address complex scenes involving variations in race, age, and individual differences. Many of these disentanglement approaches rely on auxiliary datasets or additional label information. In contrast, our work aims to disentangle expression features from non-expression attributes without the need for auxiliary data or labels, providing a

more efficient and streamlined approach to FER.

In this study, we introduce DICE-FER to overcome the limitations of existing approaches in achieving identity-invariant FER. Unlike previous methods that generate synthetic images or use subspaces to compare expressions within the same identity, our model offers a more efficient solution. DICE-FER disentangles identity from expression through mutual information (MI) estimation, eliminating the cost of subject annotation and the need for expensive image reconstruction or generation. By processing paired images with shared attributes, our model creates a feature subspace consisting of shared representations (for expression) and exclusive representations (for identity). We encode common expression representation by maximizing cross-referenced MI while minimizing the MI between expression and identity representations through an adversarial ojective. Our contibutions are summarized below:

1. We propose DICE-FER, a mutual information estimation-based approach that learns disentangled representations for identity and expression without relying on identity labels or resource-intensive image reconstruction and generation methods.

2. We introduce an innovative two-stage training methodology: initially, the expression representation is learned by estimating and maximizing cross-referenced mutual information; subsequently, identity features are learned by maximizing mutual information while minimizing the mutual information between expression and identity features.

3. We incorporate an adversarial objective to effectively minimize mutual information within expression and identity representations, overcoming the limitations of the statistical networks method described in Section 4.

4. We validate our approach experimentally on the CK+, Oulu-CASIA, RAF-DB, and AffectNet datasets, demonstrating its superiority over state-of-the-art methods.

2. Related Work

2.1. Identity-Invariant FER

Numerous methods have been developed to address the impact of identity factors on facial expression recognition (FER). Modern approaches predominantly leverage deep learning, with deep metric learning gaining prominence for its effectiveness in tasks like face recognition. For instance, Cai *et al.* [7] proposed the island loss, which minimizes intra-class variations while enhancing inter-class separations, while Liu *et al.* [27] introduced a cluster loss to improve feature discriminability by clustering positive samples and separating negatives. However, these methods often emphasize feature discriminability in the final feature space, overlooking critical aspects of the network's learning process. Huang *et. al* [16] tackled identity-related challenges using StarGAN for expression synthesis combined with deep metric learning, while Chopra *et al.* [11] employed Siamese networks with contrastive loss for identifying whether image pairs belong to the same subject. Similarly, Meng *et al.* [30] proposed an identity-aware CNN to disentangle expression and identity features through an auxiliary layer, revealing that these features remain partially inseparable. Li *et al.* [22] extended this idea by creating image pairs to facilitate feature disentanglement via a Siamese network.

GAN-based methods have also been widely adopted for FER. Yang *et al.* [42] developed IA-gen, which uses conditional GANs to generate standard expressions for each identity and compares features across the same identity subspace. Other GAN-based techniques, like hard negative generation networks [26], produce identity-specific images for training radial metric learning models to disentangle identity from expressions. However, the success of these methods relies heavily on the quality of generated images, and artifacts often degrade performance. Some researchers attempt to disentangle identity and expression components directly. For example, Xie *et al.* [40] proposed a multi-path CNN leveraging autoencoder features, while Bai *et al.* [4] and Yang *et al.* [41] utilized neutral-expression images to compare query and neutral samples for expression disentanglement. These methods, however, depend on the availability of neutral images. Approach like conditional GAN [8], which average expressions across identities, sacrifice data diversity and complexity, limiting their generalizability and exchange-GAN [43] use partial feature swapping to isolate identity and expression features but face significant computational overhead, limiting their practicality for real-world applications. Recent advancements, such as TDGAN [39] and TERGAN [1], use dual encoders to independently extract identity and expression features, while methods like Cross-VAE [38], IPD-FER [17] and AGILE [2] rely on generative models to disentangle these aspects. Despite their innovations, these techniques often struggle with spontaneous, in-the-wild datasets due to variability in expressions, poses, and external factors like head movement.

While these strategies aim to address FER challenges using unsupervised or semi-supervised generative models, they often face limitations in interpretability, control, or representation quality. Additionally, their reliance on auxiliary datasets or extra label information presents further challenges. In contrast, our work focuses on disentangling expression features from non-expression confounders, such as identity, using mutual information estimation. This approach eliminates the need for additional label information, or costly image reconstruction, providing a more efficient and streamlined solution for FER.

3. Theoretical Base

This section elaborates on the mathematical foundations of information bottleneck theory and mutual information estimation in the context of deep learning [35], [15].

A. Mutual Information: Consider two random variables $M \in \mathcal{M}$ and $Z \in \mathcal{Z}$. Let $p(m, z)$ represent the joint probability density function of M and Z, with $p(m)$ and $p(z)$ being their respective marginal probability density functions. The mutual information between M and Z is given by:

$$\mathcal{I}(M, Z) = \int_{\mathcal{M}} \int_{\mathcal{Z}} p(m, z) \log \left(\frac{p(m, z)}{p(m)p(z)} \right) dm\, dz. \tag{1}$$

From this expression, $\mathcal{I}(M, Z)$ resembles the KL divergence between P_{MZ} and $P_M P_Z$, that is, $\mathcal{I}(M, Z) = D_{KL}(P_{MZ} \| P_M P_Z)$. In our approach, we draw inspiration from MINE [5] and [33] and employ the Donsker-Varadhan representation to estimate mutual information, expressed as:

$$\mathcal{I}_{DV}(M, Z) = D_{DV}(P_{MZ} \| P_M P_Z). \tag{2}$$

This method is preferred due to its robustness and effectiveness in maximizing mutual information. The mutual information estimator in this context is defined as:

$$\hat{\mathcal{I}}_{DV,\theta}(M, Z) = \mathbb{E}_{p(m,z)}[U_\theta(m, z)] - \log \mathbb{E}_{p(m)p(z)}[e^{U_\theta(m,z)}], \tag{3}$$

where $U_\theta : \mathcal{M} \times \mathcal{Z} \to \mathbb{R}$ is a deep neural network, known as the statistics network. Belghazi *et al.* [5] propose this estimator to maximize the mutual information between an image $M \in \mathcal{M}$ and its feature representation $Z \in \mathcal{Z}$, referred to as global mutual information. This feature representation Z is extracted by a deep neural network $\mathcal{E}_\psi : \mathcal{M} \to \mathcal{Z}$, and the objective function is:

$$\mathcal{L}^{\text{global}}_{\theta,\psi}(M, Z) = \hat{\mathcal{I}}_{DV,\theta}(M, Z). \tag{4}$$

Furthermore, local mutual information maximization is also suggested, which is defined by the equation:

$$\mathcal{L}^{\text{local}}_{\phi,\psi}(M, Z) = \sum_i \hat{\mathcal{I}}_{DV,\phi}(\mathcal{F}^{(i)}_\psi(M), Z). \tag{5}$$

where $\mathcal{F}_\psi(M)$ represents the information content of the spatial regions of M, and Z denotes the feature representation.

4. Proposed Method

Let M and N be two images from the domains $\mathcal{M}$ and $\mathcal{N}$, respectively, with $T_M \in \mathcal{T}_\mathcal{M}$ and $T_N \in \mathcal{T}_\mathcal{N}$ denoting their corresponding representations. These representations are divided in two components: expression components, E_M and E_N, that capture the shared representation between M and N, and identity components, I_M and I_N, that represent the identity information specific to each image. Thus, the representation for M is written as $T_M = [E_M, I_M]$, and similarly, for N, it is $T_N = [E_N, I_N]$.

To facilitate the disentangling of these representations, we introduce a two-step training process. In the first stage, the expression information between images is learned to form an expression representation (refer to Section 4.1). With the expression information established, the second stage focuses on learning the identity representation, which captures the unique subject details not present in the expression representation (see Section 4.2). An overview of the model is depicted in Figure 2.

4.1. Expression Representation Learning

In order to learn the expression encodings E_M and E_N from the images M and N, respectively, we define the encoder functions $\mathcal{E}^{exp}_{\psi_M} : \mathcal{M} \to \mathfrak{E}_\mathcal{M}$ & $\mathcal{E}^{exp}_{\psi_N} : \mathcal{N} \to \mathfrak{E}_\mathcal{N}$. The mutual information linked with the images and their expression features is estimated and optimized using the global statistics networks $U^{exp}_{\theta_M}$ and $U^{exp}_{\theta_N}$, as well as the local statistics networks $U^{exp}_{\phi_M}$ and $U^{exp}_{\phi_N}$. Equation (4) and (5) are employed for this purpose. Unlike MINE [5], we utilize a technique known as swapping the shared (expression) representations to compute the cross-referenced mutual information across images M and N. This approach considers the fixed coefficients μ^{exp} and ν^{exp}, which evaluate the relative importance of the global and local mutual information components. An essential component of the suggested method is the exchange of expression representations, which guarantees the removal of exclusive (identity) information from each image (see to Section 5.5).

$$\begin{aligned} \mathcal{L}^{exp}_{MI} = \mu^{exp} &\left(\mathcal{L}^{\text{global}}_{\theta_M,\psi_N}(M, E_N) + \mathcal{L}^{\text{global}}_{\theta_N,\psi_M}(N, E_M) \right) \\ &+ \nu^{exp} \left(\mathcal{L}^{\text{local}}_{\phi_M,\psi_N}(M, E_N) + \mathcal{L}^{\text{local}}_{\phi_N,\psi_M}(N, E_M) \right) \end{aligned} \tag{6}$$

Also, it is necessary for images M and N to possess same expression representations, thus $E_M = E_N$. For this, the L_1 distance is minimized between the expression representations E_M and E_N in the following manner:

$$L_1 = \mathbb{E}_{p(E_m,E_n)}[|E_M - E_N|]. \tag{7}$$

Consequently, expression learning objective is a linear combination of Eq. 6 and 7, where δ is a weighting parameter. This is given by:

$$\max_{\{\psi,\theta,\phi\}_{M,N}} \mathcal{L}^{\text{exp}} = \mathcal{L}^{exp}_{MI} - \delta L_1. \tag{8}$$

4.2. Identiy Representation Learning

By now, the model has the ability to effectively extract the expression representations E_M and E_N. Define the encoder functions $\mathcal{E}^{id}_{\eta_M} : \mathcal{M} \to \mathfrak{I}_\mathcal{M}$ and $\mathcal{E}^{id}_{\eta_N} : \mathcal{N} \to \mathfrak{I}_\mathcal{N}$ as necessary for extracting the identity representations I_M

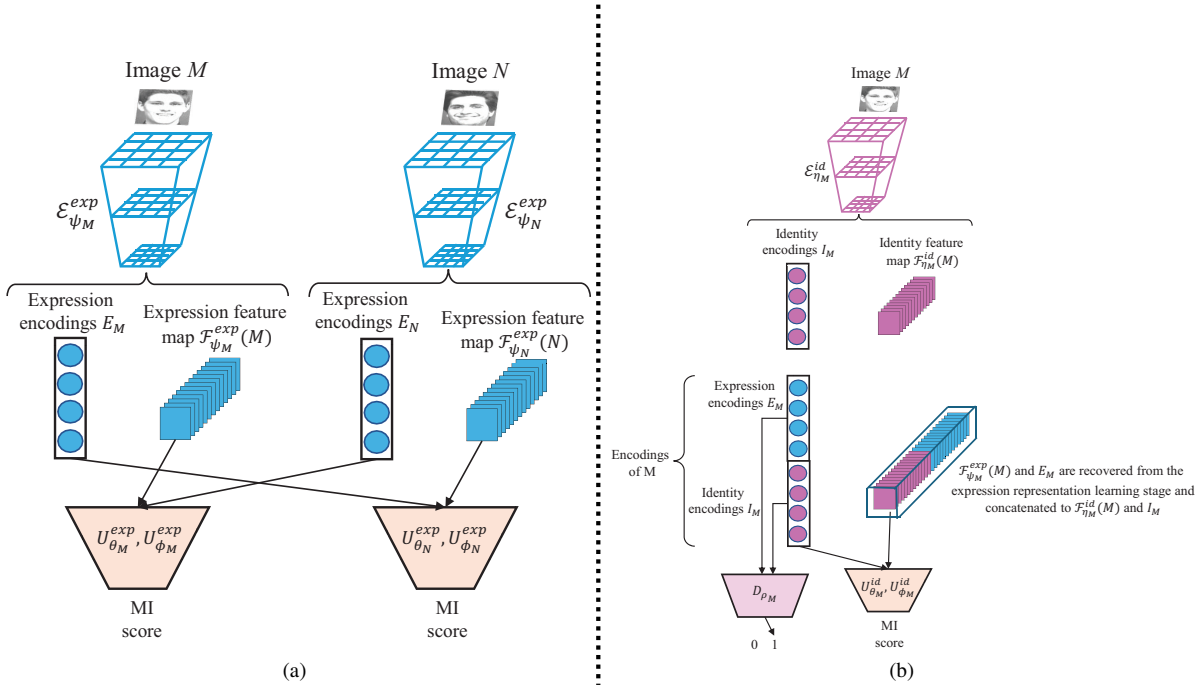


Fig. 2. Overall framework. (a) The initial step involves learning the expression representation. Images M and N are processed through expression representation encoders to extract E_M and E_N. MI maximization is performed via statistics networks for M and E_N, and N and E_M. (b) The subsequent step focuses on learning the identity representation. Image M is processed through identity encoder to acquire I_M. The statistics networks maximize MI between M and $T_M = [E_M, I_M]$, whereas the MI between E_M and I_M is minimized by the discriminator. This procedure is similarly applied to image N to obtain I_N (Best vsible in colour).

and I_N from images M and N, respectively. We estimate and maximize the mutual information between the image M and its corresponding representation T_M to learn these representations. This representation includes both expression and identity features, denoted as $T_M = [E_M, I_M]$. The identical process is executed between the image N and $T_N = [E_N, I_N]$, as illustrated in Eq. (9), where μ^{id} and ν^{id} are fixed weighting parameters. We compute the mutual information using the global statistical network $U^{id}_{\theta_M}$ and $U^{id}_{\theta_N}$, as well as the local statistical networks $U^{id}_{\phi_M}$ and $U^{id}_{\phi_N}$. By maintaining a constant expression representation, we ensure that the identity representation incorporates the information that is unique to the image and is not included in the expression representation.

$$\begin{aligned}\mathcal{L}^{id}_{MI} = \mu^{id}\left(\mathcal{L}^{\text{global}}_{\theta_M,\eta_M}(M,T_M) + \mathcal{L}^{\text{global}}_{\theta_N,\eta_N}(N,T_N)\right) \\ + \nu^{id}\left(\mathcal{L}^{\text{local}}_{\phi_M,\eta_M}(M,T_M) + \mathcal{L}^{\text{local}}_{\phi_N,\eta_N}(N,T_N)\right)\end{aligned} \tag{9}$$

Conversely, it is required that the representation I_M excludes any information captured by E_M while ensuring that the mutual information between M and T_M is maximized. Consequently, it is imperative to minimize the mutual information between I_M and T_M. Although mutual information estimation and maximization via Eq. (3) effectively enhance mutual information, the use of statistics networks encounters convergence problems during the calculation and minimization of MI. Minimizing Eq. (3) causes statistics networks to diverge. To circumvent this we minimize the mutual information between E_M and I_M (i.e., $\mathcal{I}(E_M, I_M)$) through an adversarial objective as illustrated in Eq. (10). Minimizing $\mathcal{I}(E_M, I_M)$ is tantamount to minimizing $D_{DV}(P_{E_M I_M} \| P_{E_M} P_{I_M})$, which is accomplished via a adversarial framework. Consequently, a discriminator network D_{ρ_M}, parameterized by ρ_M, is trained to differentiate between representations sampled from $P_{E_M I_M}$ and $P_{E_M} P_{I_M}$, treating the former as fake and the latter as real samples. To obtain samples from $P_{E_M I_M}$, the image M is passed through the encoders $\mathcal{E}^{exp}_{\psi_M}$ and $\mathcal{E}^{id}_{\eta_M}$ to extract (E_M, I_M). Conversely, samples from $P_{E_M} P_{I_M}$ are obtained by shuffling the identity representations within a batch. The encoder function $\mathcal{E}^{id}_{\eta_M}$ is designed to generate identity representations I_M such that, when combined with E_M, they mimic samples drawn from $P_{E_M} P_{I_M}$. By minimizing Equation (10), the Donsker-Varadhan divergence $D_{DV}(P_{E_M I_M} \| P_{E_M} P_{I_M})$ is minimized, leading to the minimization of mutual information between I_M and E_M. An analogous method to produce samples from the joint distribution's marginals is suggested by [6], [18]. These models utilize an adversarial objective to ensure that each dimension of the representation is independent of the others. In contrast, our approach employs an adversarial objective to distinctly separate the dimensions of the expression repre-

sentation from identity.

$$\mathcal{L}_M^{\text{adv}} = \mathbb{E}_{p(e_M)p(i_M)}[\log D_{\rho_M}(E_M, I_M)] + \mathbb{E}_{p(e_M,i_M)}[\log(1 - D_{\rho_M}(E_M, I_M))] \quad (10)$$

Identity learning objective is a linear formulation of preceding terms, where ζ represents a weighting parameter. This is given by:

$$\max_{\{\eta,\theta,\phi\}_{M,N}} \min_{\{\rho\}_{M,N}} \mathcal{L}^{id} = \mathcal{L}_{MI}^{id} - \zeta^{\text{adv}}(\mathcal{L}_M^{\text{adv}} + \mathcal{L}_N^{\text{adv}}) \quad (11)$$

Finally, a classifier trained on pure expression representation space (E_M/E_N) is used for the FER task.

5. Experiments

5.1. Datasets

5.1.1 CK+

The CK+ dataset [28] is a laboratory-controlled collection of posed facial expressions, capturing data from 210 individuals across diverse age groups, genders, and regions. It comprises 593 image sequences from 123 subjects, aged 18 to 50 years. The grayscale images, sized 640×490 pixels, depict transitions from neutral expressions to peak emotions. For evaluation, we utilized the first and last frames of each sequence, creating a subset of 500 images representing the seven basic expressions.

5.1.2 Oulu-CASIA

The Oulu-CASIA dataset [48] is a prominent benchmark used for facial expression recognition (FER) and face verification studies. Collected in a controlled laboratory setting, it includes six basic expressions from 80 participants aged 23 to 58. The dataset features 480 sequences, with each individual contributing one sequence per expression. Each sequence begins with a neutral expression and transitions to a peak expression. Following standard research protocols [47], the last three frames of each sequence were chosen for model training and evaluation.

5.1.3 RAF-DB

RAF-DB [20] is a large-scale, in-the-wild facial expression dataset comprising approximately 30,000 diverse facial images collected from thousands of individuals online. The labeling process involved 315 human coders, with final annotations determined using crowdsourcing methods. Each image was reviewed by around 40 independent labelers to ensure annotation reliability. The dataset includes 12,271 training samples and 3,068 test samples categorized into seven basic emotions: angry, disgust, fear, happy, neutral, sad, and surprise. While RAF-DB also provides compound expressions labeled into 11 classes, these were not utilized in our experiments.

5.1.4 AffectNet

The AffectNet database [31] is a large-scale collection comprising over 400,000 labeled images, making it the most extensive FER dataset available. These images were sourced from the internet using three different search engines and keywords related to facial expressions. For our experiments, we focus on the seven basic expression categories, similar to RAF-DB, which provides approximately 280,000 images for training. The validation set includes 500 samples per category, totaling 3,500 images.

5.2. Preprocessing

The datasets are preprocessed using MTCNN [47] to extract facial landmarks. A similarity transformation crops and aligns the primary facial region, followed by whitening to reduce redundancy. Images are resized to 112×112×1. To counter the scarcity of expression images, we applied 14 data augmentation techniques, including rotations (±15°, ±10°, ±5°, and 0°) and horizontal flipping. A 10-fold cross-validation is applied to all three datasets. An image pair is created by sampling two images of the same expression from each dataset. We train our model on all datasets to learn a shared (expression) and unique (identity) 64-dimensional representation.

5.3. Implementation details

The architectural details of the model consist of ResNet-18 [14] based encoders pre-trained on CASIA-WebFace [44], with the statistics networks derived from those used in MINE [5]. The discriminator is designed as a network comprising three fully connected layers. For classifier, two fully-connected hidden layers are used. We initialize and train each network using a batch size of 32. We used 100 epochs for training our model. Optimization is done using the Adam optimizer [19] with a learning rate of 10^{-4}. The loss coefficients applied in the model are $\mu^{exp} = \mu^{id} = 0.5$, $\nu^{exp} = \nu^{id} = 1.0$, and $\delta = 0.1$. The effect of coefficient ζ^{adv} is discussed in detail in the ablation stud. The computational platform consists of a Ryzen 9 octa-core processor with 16GB RAM and a 4GB NVIDIA GeForce RTX-3050 graphics card, with PyTorch used as the deep learning framework.

5.4. Quantitative evaluation of disentanglement

We quantitatively evaluate disentanglement performance using a modified Mutual Information Gap (MIG) metric, adapted for unsupervised settings. Unlike traditional MIG, which relies on ground-truth labels, we compute the gap between (1) mutual information (MI) of expression features E_M/E_N (maximized across paired images) and (2) MI between expression (E_M) and identity (I_M) features (minimized via adversarial training), i.e., MIG $= \mathcal{I}(E_M, E_N) -$

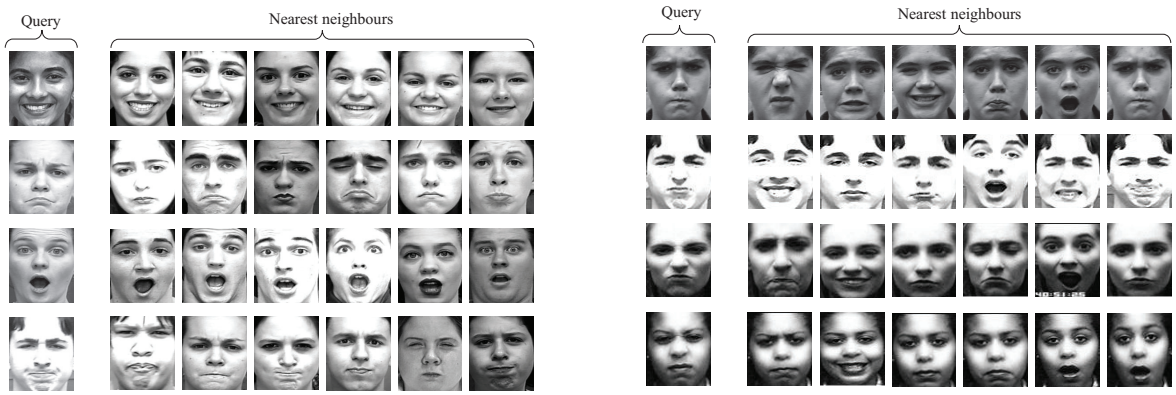


Fig. 3. Image retrieval on the CK+ dataset. Retrieved images using the (a) shared (Expression) representations (on the left) and the (b) exclusive (Idnetity) representations (on the right).

TABLE 1
MODEL EVALUATION FOR EXPRESSION DISENTANGLEMENT ON FOUR DATASETS

Methods	MIG (↑)			
	CK+	Oulu-CASIA	RAF-DB	AffectNet
TDGAN [39]	0.454	0.400	0.365	0.350
Cross-VAE [38]	0.490	0.411	0.371	0.383
TERGAN [1]	0.496	0.432	0.388	0.388
IPD-FER [17]	0.465	0.455	0.394	0.427
AGILE [2]	0.588	0.481	0.400	0.430
DICE-FER (ours)	**0.592**	**0.497**	**0.418**	**0.448**

TABLE 2
ABLATION STUDY: EFFECT OF EACH ELEMENT ON ACCURACY ACROSS THREE DATASETS

Feature/Element	Accuracy (%)			
	CK+	Oulu-CASIA	RAF-DB	AffectNet
Ideal E_M	100.00	100.00	100.00	100.00
Baseline	**99.50**	**91.10**	**90.30**	**64.36**
Non-SER	93.45	80.50	82.37	52.95
$\delta = 0$	98.75	88.50	87.40	60.41
$\mu^{exp} = 0$	97.30	85.60	84.89	60.06
$\nu^{exp} = 0$	96.80	83.00	83.20	58.80

$\mathcal{I}(E_M, I_M)$. Higher MIG scores reflect stronger separation between expression and identity features. As shown in Table 1, DICE-FER consistently outperforms prior methods (TDGAN [39], Cross-VAE [38], IPD-FER [17], AGILE [2]) across all datasets. Notably, on challenging in-the-wild datasets like RAF-DB and AffectNet, DICE-FER achieves 0.418 and 0.448 MIG (vs. TDGAN's 0.365/0.350), demonstrating robust retention of expression cues despite identity, pose, and background variability.

5.5. Qualitative disentanglement through image retrieval

In the context of FER, image retrieval serves as a critical qualitative tool to validate disentanglement between expression (shared representation) and identity (exclusive representation). For a query image M, its shared representation E_M (encoding expression attributes) is compared to shared representations of all other images in the dataset using a cosine similarity. The top-k closest matches are retrieved. These should share the same shared attributes (expression) but vary in exclusive attributes (identity). Similarly, the exclusive representation I_M (encoding identity attributes) is compared to exclusive representations of other images. The

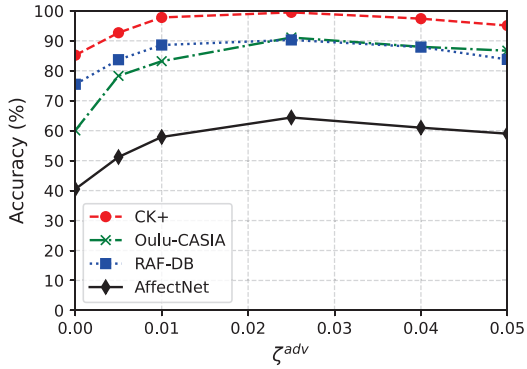
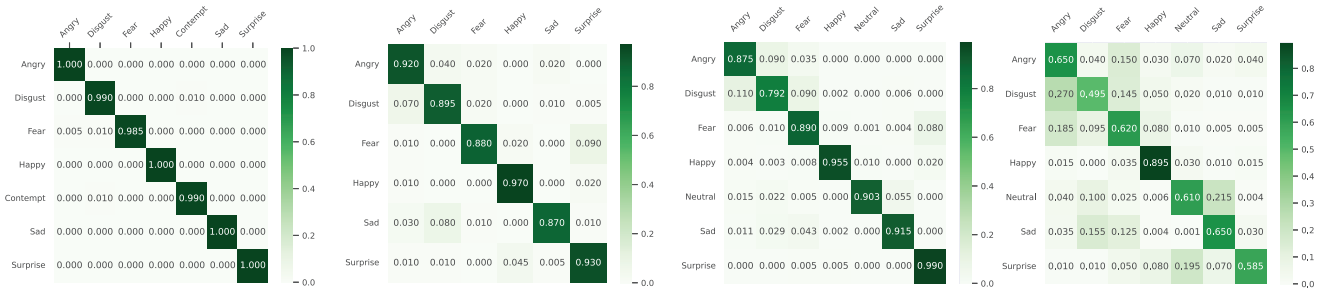


Fig. 4. Various values of ζ^{adv} are employed to learn the expression representation, with the results plotted to show accuracy as a function of ζ^{adv}.

retrieved images should share the same exclusive attributes (identity) but vary in shared attributes (expression). These are evident from the results shown in Figure 3 on CK+ dataset, where querying with the shared representation of expression "Happiness" retrieves other "happy faces" with different identities, while querying with the exclusive representation of a "particular identity" retrieves expressions of different classes with same identity.



(a) CK+ (b) Oulu-CASIA (c) RAF-DB (d) AffectNet

Fig. 5. Plots of confusion matrix on three datasets (a) CK+, (b) Oulu-CASIA, (c) RAF-DB, and (d) AffectNet

TABLE 3
PERFORMANCE COMPARISON ON THE CK+ DATASET.

Methods	Setting	Expression	Metric			
			Accuracy (%)	Precision	Recall	F1 Score
DeRL(2018) [41]	image-based	7	97.30	0.98	0.97	0.975
IA-gen (2018) [42]	image-based	7	96.57	0.96	0.98	0.969
ADFL (2019) [4]	image-based	7	98.17	0.99	0.99	0.990
DDL (2020) [32]	image-based	7	99.16	0.99	0.97	0.979
TDGAN (2020) [39]	image-based	7	97.53	0.98	0.98	0.980
Cross-VAE (2020) [38]	image-based	7	94.96	0.95	0.93	0.939
IE-DBN (2021) [46]	image-based	7	96.02	0.97	0.94	0.954
TER-GAN (2021) [1]	image-based	7	98.47	0.99	0.97	0.979
Huang *et. al* (2021) [16]	image-based	7	98.65	0.98	0.99	0.985
IF-GAN (2021) [8]	image-based	6	97.52	0.96	0.96	0.960
HSIC-Contrastive loss (2022) [22]	image-based	7	98.60	0.98	0.97	0.975
IPD-FER (2022) [17]	image-based	7	98.65	0.97	0.97	0.970
DR-ALNet (2024) [10]	image-based	7	99.34	0.99	0.99	0.990
AGILE (2024) [2]	image-based	7	99.00	1.00	0.99	**0.995**
ResNet-18	image-based	7	97.39	0.95	0.96	0.955
DICE-FER (ours)	image-based	7	**99.50**	**1.00**	**0.99**	**0.995**

TABLE 4
PERFORMANCE COMPARISON ON THE OULU-CASIA DATASET IN TERMS OF 6 EXPRESSIONS (WITHOUT NEUTRAL)

Methods	Setting	Metric			
		Accuracy (%)	Precision	Recall	F1 Score
DeRL (2018) [41]	image-based	88.00	0.89	0.86	0.874
IA-gen (2018) [42]	image-based	88.92	0.87	0.88	0.875
ADFL (2019) [4]	image-based	87.50	0.86	0.88	0.869
Exchange GAN (2020) [43]	sequence-based	86.33	0.87	0.88	0.875
DDL (2020) [32]	image-based	88.26	0.89	0.88	0.885
Cross-VAE (2020) [38]	image-based	86.87	0.88	0.86	0.869
IE-DBN (2021) [46]	image-based	85.21	0.85	0.85	0.850
Huang *et. al* (2021) [16]	image-based	87.90	0.88	0.88	0.880
HSIC-Contrastive loss (2022) [22]	image-based	88.82	0.87	0.88	0.875
DR-ALNet (2024) [10]	image-based	90.00	0.91	0.90	0.905
AGILE (2024) [2]	image-based	90.00	0.91	0.89	0.899
ResNet-18	image-based	86.00	0.85	0.86	0.855
DICE-FER (ours)	image-based	**91.10**	**0.92**	**0.92**	**0.920**

TABLE 5
PERFORMANCE COMPARISON ON RAF-DB DATASET

Methods	Metric			
	Accuracy (%)	Precision	Recall	F1 Score
DDL (2020) [32]	87.71	0.88	0.85	0.864
Cross-VAE (2020) [38]	84.81	0.85	0.85	0.850
IF-GAN (2021) [8]	88.33	0.89	0.90	0.895
TDGAN (2020) [39]	81.91	0.81	0.80	0.805
IE-DBN (2021) [46]	84.75	0.85	0.83	0.839
IPD-FER (2022) [17]	88.89	0.89	0.89	0.890
DR-ALNet (2024) [10]	89.34	0.90	0.88	0889
ResNet-18	87.48	0.85	0.87	0.859
DICE-FER (ours)	**90.30**	**0.91**	**0.90**	**0.905**

TABLE 6
PERFORMANCE COMPARISON ON AFFECTNET DATASET

Methods	Metric			
	Accuracy (%)	Precision	Recall	F1 Score
PG-CNN (2018) [25]	55.33	0.56	0.54	0.549
Separate loss (2019) [23]	58.89	0.60	0.58	0.590
IPA2LT (2018) [45]	57.31	0.57	0.57	0.570
RAN (2020) [37]	59.50	0.60	0.59	0.595
SNA (2020) [12]	62.70	0.62	0.61	0.615
BregNet (2019) [13]	63.54	**0.65**	0.64	0.645
Chen *et al.* (2021) [9]	61.98	0.62	0.60	0.610
IPD-FER (2022) [17]	62.23	0.63	0.61	0.620
THIN (2022) [3]	63.97	**0.65**	0.62	0.634
ResNet-18	59.06	0.59	0.59	0.590
DICE-FER (ours)	**64.36**	**0.65**	**0.65**	**0.650**

5.6. Ablation Study

To assess each element's contribution during expression representation learning, we remove them and evaluate the impact on classification accuracy. Our baseline settings (Section 5.3) are: $\mu^{exp} = 0.5$, $\nu^{exp} = 1.0$, $\delta = 0.1$, and swapped expression representations (SER). Ablation results for the three datasets (Tables 2) show that swapping expression representations is key to disentanglement. Without SER, accuracy on expression representation decreases, indicating expression representations capture identity information. Removing the L1 distance ($\delta = 0$) reduces the accuracy slightly, and eliminating global mutual information ($\mu^{exp} = 0$) causes slight mix-up of identity with expression features. Local mutual information ($\nu^{exp} = 0$) is crucial for capturing expression information, with accuracy dropping sharply when excluded. These findings affirm that all loss terms contribute to the overall performance of the model.

1) Impact of parameter ζ^{adv}: To understand the impact of the parameter ζ^{adv} on minimizing mutual information between expression and identity representations, we trained our model using different values of $\zeta^{\text{adv}} \in \{0.0, 0.005, 0.010, 0.025, 0.04, 0.05\}$. The expression representations were then used for classification tasks. The results, shown in Fig. 4, indicate that lower ζ^{adv} is linked to weak disentanglement causing expression features to

TABLE 7
CROSS-DATASET EXPERIMENTS

Method	Train	Test	Accuracy (%)
MSDModel [34]	CK+	Oulu-CASIA	45.83
	Oulu-CASIA	CK+	55.55
gACNN [24]	RAF-DB	CK+	81.07
SPWFA-SE [21]	RAF-DB	CK+	81.72
	AffectNet	CK+	85.44
VTFF-FER [29]	RAF-DB	CK+	81.88
	AffectNet	CK+	86.24
AGILE [2]	CK+	Oulu-CASIA	48.12
	Oulu-CASIA	CK+	61.32
DICE-FER (ours)	CK+	Oulu-CASIA	**50.67**
	Oulu-CASIA	CK+	**62.48**
	RAF-DB	CK+	**82.34**
	AffectNet	CK+	**86.91**

contain identity information, degrading classification performance and compromising the purity of the expression space. Optimal value of ζ^{adv} achieves a balanced disentanglement, with pure expression space, leading to high accuracy for the FER task. Higher ζ^{adv} over-enforces disentanglement, stripping away useful expression information, which can degrade classification accuracy on FER task.

5.7. Comparison with SOTA methods

Tables 3–6 highlight DICE-FER's superiority across datasets. On CK+ (Table 3), it achieves near-perfect 99.50% accuracy (vs. IA-GAN's 98.17%) and flawless precision (1.00), excelling in controlled settings by isolating subtle expression distinctions. For Oulu-CASIA (Table 4), 91.10% accuracy (vs. DR-ALNet's 90.60%) reflects robustness to occlusions like glasses, while on RAF-DB (Table 5), 90.30% accuracy (vs. IPD-FER's 88.89%) underscores effectiveness in real-world clutter. AffectNet results (Table 6: 64.36% vs. BregNet's 63.54%) validate scalability to noisy, large-scale data. Confusion matrices (Fig. 5) reveal errors align with human perceptual ambiguities—e.g., fear vs. disgust or neutral vs. sad—highlighting challenges in fine-grained distinctions. The adversarial mutual information minimization ensures identity features do not corrupt expression representations, yielding balanced precision-recall trade-offs (F1: 0.99 on CK+). By decoupling static identity attributes from dynamic expressions, DICE-FER advances identity-invariant recognition, offering a scalable, annotation-free solution for both controlled and in-the-wild scenarios.

5.8. Cross-Dataset Evaluation

DICE-FER demonstrates robust generalization across diverse domains, as evidenced by cross-dataset experiments in Table 7. When trained on CK+ (controlled, posed expressions) and tested on Oulu-CASIA (occlusions, demographic diversity), it achieves 50.67% accuracy, outperforming MSDModel (45.83%) and AGILE (48.12%). Conversely, training on AffectNet (in-the-wild) and testing on CK+ yields 86.91% accuracy, surpassing SPWFA-SE (85.44%) and VTFF-FER (86.24%). These gains stem from DICE-FER's disentangled representations, which suppress identity-specific biases (e.g., facial structure) and prioritize expression semantics, enabling adaptability to unseen domains with varying noise levels and demographics.

6. Conclusion

This paper presents DICE-FER, a novel approach for identity-invariant FER. Instead of generating synthetic images or using subspaces for comparison, our model disentangles identity from expression through mutual information estimation. Our method eliminates the costly need for subject annotation and image reconstruction. By processing paired images with shared attributes, we create shared (expression) and exclusive (identity) representations. We then maximize mutual information between images and their expression representations and use an adversarial framework to minimize the overlap between these representation types. Our method demonstrates improved performance on the CK+, Oulu-CASIA, RAF-DB, and AffectNet datasets, outperforming state-of-the-art techniques.

References

[1] Kamran Ali and Charles E Hughes. Facial expression recognition by using a disentangled identity-invariant expression representation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9460–9467. IEEE, 2021. 2, 6, 7

[2] Mohd Aquib, Nishchal K Verma, and M Jaleel Akhtar. Agile: Attribute-guided identity independent learning for facial expression recognition. *IEEE Transactions on Affective Computing*, (01):1–16, 2024. 2, 6, 7, 8

[3] Estèphe Arnaud, Arnaud Dapogny, and Kevin Bailly. Thin: Throwable information networks and application for facial expression recognition in the wild. *IEEE Transactions on Affective Computing*, 14(3):2336–2348, 2022. 8

[4] Mengchao Bai, Weicheng Xie, and Linlin Shen. Disentangled feature based adversarial learning for facial expression recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 31–35. IEEE, 2019. 1, 2, 7

[5] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018. 3, 5

[6] Philemon Brakel and Yoshua Bengio. Learning independent features with adversarial nets for non-linear ica. *arXiv preprint arXiv:1710.05050*, 2017. 5

[7] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 302–309. IEEE, 2018. 2

[8] Jie Cai, Zibo Meng, Ahmed Shehab Khan, James O'Reilly, Zhiyuan Li, Shizhong Han, and Yan Tong. Identity-free facial expression recognition using conditional generative adversarial network. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1344–1348. IEEE, 2021. 1, 2, 7

[9] Boyu Chen, Wenlong Guan, Peixia Li, Naoki Ikeda, Kosuke Hirasawa, and Huchuan Lu. Residual multi-task learning for facial landmark localization and expression recognition. *Pattern Recognition*, 115:107893, 2021. 8

[10] Puhua Chen, Zhe Wang, Shasha Mao, Xinyue Hui, and Yanning Hu. Dual-branch residual disentangled adversarial learning network for facial expression recognition. *IEEE Signal Processing Letters*, 2024. 2, 7

[11] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005. 2

[12] Yongjian Fu, Xintian Wu, Xi Li, Zhijie Pan, and Daxin Luo. Semantic neighborhood-aware deep facial expression recognition. *IEEE Transactions on Image Processing*, 29:6535–6548, 2020. 8

[13] Behzad Hasani, Pooran Singh Negi, and Mohammad H Mahoor. Bounded residual gradient networks (breg-net) for facial affect computing. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019. 8

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[15] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 3

[16] Wei Huang, Siyuan Zhang, Peng Zhang, Yufei Zha, Yuming Fang, and Yanning Zhang. Identity-aware facial expression recognition via deep metric learning based on synthesized images. *IEEE Transactions on Multimedia*, 24:3327–3339, 2021. 2, 7

[17] Jing Jiang and Weihong Deng. Disentangling identity and pose for facial expression recognition. *IEEE Transactions on Affective Computing*, 13(4):1868–1878, 2022. 2, 6, 7, 8

[18] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658. PMLR, 2018. 5

[19] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[20] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 5

[21] Yingjian Li, Guangming Lu, Jinxing Li, Zheng Zhang, and David Zhang. Facial expression recognition in the wild using multi-level features and attention mechanisms. *IEEE Transactions on Affective Computing*, 14(1):451–462, 2020. 8

[22] Yande Li, Yonggang Lu, Minglun Gong, Li Liu, and Ligang Zhao. Dual-channel feature disentanglement for identity-invariant facial expression recognition. *Information Sciences*, 608:410–423, 2022. 2, 7

[23] Yingjian Li, Yao Lu, Jinxing Li, and Guangming Lu. Separate loss for basic and compound facial expression recognition in the wild. In *Asian conference on machine learning*, pages 897–911. PMLR, 2019. 8

[24] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018. 8

[25] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Patch-gated cnn for occlusion-aware facial expression recognition. In *2018 24th international conference on pattern recognition (ICPR)*, pages 2209–2214. IEEE, 2018. 8

[26] Xiaofeng Liu, BVK Vijaya Kumar, Ping Jia, and Jane You. Hard negative generation for identity-disentangled facial expression recognition. *Pattern Recognition*, 88:1–12, 2019. 2

[27] Xiaofeng Liu, BVK Vijaya Kumar, Jane You, and Ping Jia. Adaptive deep metric learning for identity-aware facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 20–29, 2017. 2

[28] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010. 5

[29] Fuyan Ma, Bin Sun, and Shutao Li. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, 14(2):1236–1248, 2021. 8

[30] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. Identity-aware convolutional neural network for facial expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 558–565. IEEE, 2017. 2

[31] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 5

[32] Delian Ruan, Yan Yan, Si Chen, Jing-Hao Xue, and Hanzi Wang. Deep disturbance-disentangled learning for facial expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2833–2841, 2020. 1, 7

[33] Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortner. Learning disentangled representations via mutual information estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 205–221. Springer, 2020. 3

[34] Xiao Sun, Pingping Xia, and Fuji Ren. Multi-attention based deep neural network with hybrid features for dynamic sequential facial expression recognition. *Neurocomputing*, 444:378–389, 2021. 8

[35] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. Ieee, 2015. 3

[36] Michel F Valstar, Marc Mehu, Bihan Jiang, Maja Pantic, and Klaus Scherer. Meta-analysis of the first facial expression recognition challenge. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):966–979, 2012. 1

[37] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020. 8

[38] Haozhe Wu, Jia Jia, Lingxi Xie, Guojun Qi, Yuanchun Shi, and Qi Tian. Cross-vae: Towards disentangling expression from identity for human faces. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4087–4091. IEEE, 2020. 1, 2, 6, 7

[39] Siyue Xie, Haifeng Hu, and Yizhen Chen. Facial expression recognition with two-branch disentangled generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6):2359–2371, 2020. 1, 2, 6, 7

[40] Siyue Xie, Haifeng Hu, and Yongbo Wu. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. *Pattern recognition*, 92:177–191, 2019. 2

[41] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2168–2177, 2018. 1, 2, 7

[42] Huiyuan Yang, Zheng Zhang, and Lijun Yin. Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 294–301. IEEE, 2018. 1, 2, 7

[43] Lie Yang, Yong Tian, Yonghao Song, Nachuan Yang, Ke Ma, and Longhan Xie. A novel feature separation model exchange-gan for facial expression recognition. *Knowledge-Based Systems*, 204:106217, 2020. 2, 7

[44] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 5

[45] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237, 2018. 8

[46] Haifeng Zhang, Wen Su, Jun Yu, and Zengfu Wang. Identity–expression dual branch network for facial expression recognition. *IEEE transactions on cognitive and developmental systems*, 13(4):898–911, 2020. 7

[47] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 5

[48] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and vision computing*, 29(9):607–619, 2011. 5